BD Single-Cell Multiomics

Bioinformatics Handbook

23-21713(04)
2023-01

For Research Use Only

Regulatory information

For Research Use Only. Not for use in diagnostic or therapeutic procedures.

History

Revision	Date	Change made
Doc ID: 54169 Rev. 1.0	2017-09	Initial release.
Doc ID: 54169 Rev. 2.0	2017-11	—Added content on sample multiplexing. See Step 7. Determine the sample of origin (sample multiplexing only) on page 20 and Reviewing sequencing analysis output files on page 34. —Added content on specifying gene targets.
Doc ID: 54169 Rev. 3.0	2018-01	—Updated content of BD Data View to v1.1, which includes these new features: New color options for plots, highlight selected annotated groups in plots, filter data table by cells based on gene expression, new calculation on fold changes and mean gene expression, new option to modify data table names and annotation list names. —Added two examples for use with BD Data View. —Expanded information on selecting a transcript, selecting primers, and output files.
Doc ID: 54169 Rev. 4.0	2018-04	—Added another example for use with BD Data View.
Doc ID: 54169 Rev. 5.0	2018-07	—Added metrics outputs for BD® AbSeq. See BD Rhapsody™ sequencing analysis on page 6. —Updated to BD Data View v1.2. Some new features include: –Automatic detection of AbSeq markers in data tables –New Gene A v. Gene B feature to compare two gene markers –Combine one or more data tables –Differential expression of >1,500 genes

Revision	Date	Change made
Doc ID: 54169 Rev. 6.0	2018-10	—Updated cross references from system user guides to instrument user guides. —Changed content to say that a BAM file is sorted according to the alignment coordinates of R2 reads on each chromosome. See BAM and BAM Index on page 45. —Added recommendation to analyze datasets that are ≤1TB in size. See Understanding the BD Rhapsody™ Analysis pipeline step-by-step (page 11). —Updated output file name in example to Combined_<sample_multiplex_name>_DBEC_MolsPerCell.csv.
Doc ID: 54169 Rev. 7.0 23-21713-00	2019-07	—Added content for BD Rhapsody™ System Whole Transcriptome Analysis (WTA). —Updated some step parameters. —Revised recommendation to analyze datasets from ≤TB to ≤100 GB.
Doc ID: 54169 Rev. 8.0 23-21713-01	2019-10	—Added content for BD Rhapsody™ System Whole Transcriptome Analysis (WTA) and AbSeq.
Doc ID: 54169 Rev. 9.0 23-21713(02)	2021-08	—Updated BD Rhapsody™ System Whole Transcriptome Analysis (WTA) and AbSeq. —Added sequencing analysis outputs for Protein aggregates experimental, Bioproduct statistics, Sample tag metrics, Sample tag calls, and Per sample folder. —Added a new step “TCR and BCR analysis (if applicable).”
23-21713(03)	2022-08	—Added new file definitions: VDJ Dominant Contigs AIRR, VDJ Unfiltered Contigs AIRR. —Added pipeline report.
23-21713(04)	2023-01	—Added support for BD® Flex Single-Cell Multiplexing Kit. —Added support for new cell label structure for BD Rhapsody™ Enhanced Cell Capture Beads v2.0. —Added support for v1.12 Bioinformatics pipeline. —Improved VDJ outputs.

Contents

1. Introduction

About this handbook

This handbook is a comprehensive reference to help you prepare and analyze single-cell libraries with the BD Rhapsody™ Single-Cell Analysis system or the BD Rhapsody™ Express Single-Cell Analysis system. Major aspects of the BD® Single-Cell Multiomics bioinformatics workflow are covered. This reference explains the BD® Single-Cell Multiomics sequencing analysis algorithms to deepen your understanding of how single-cell mRNA and protein (AbSeq) expression profiles are generated. In addition, the handbook defines every analysis metric.

The BD Single-Cell Multiomics team

2. BD Rhapsody™ sequencing analysis

How to use this chapter

This chapter provides in-depth information on the process, output metrics, and interpretation of output from BD Rhapsody™ sequencing analysis:

Section	Information
BD Rhapsody™ sequencing analysis on page 6	Detailed description of each step in the BD Rhapsody™ pipeline
Reviewing sequencing analysis output files on page 34	Definitions of the sequencing analysis output metrics
Interpreting output metrics on page 61	Recommended solutions to possible problems during sequencing analysis

Understanding the BD Rhapsody™ Analysis pipeline step-by-step

Introduction

This section provides an in-depth description of each step in the BD Rhapsody™ Analysis pipeline.

For instructions on running the pipeline, see the *BD® Single-Cell Multiomics Analysis Setup User Guide* (23-21333).

Single-Cell Multiomics technical publications are available for download from the BD® Single-Cell Multiomics Resource Library at scomix.bd.com/hc/en-us/categories/360000838932-Resource-Library.

Overview

The BD Rhapsody™ assays are used to create sequencing libraries from single-cell multiomic experiments.

The analysis pipeline works with paired-end FASTQ R1 and R2 files. R1 reads contain information on the cell label and molecular identifier, and R2 reads contain information on the bioproduct. See Figure 1.

Figure 1 Structure of read pair that is generated by sequencing the libraries prepared with BD Rhapsody™ assays.

Pipeline overview

After sequencing, the targeted analysis pipeline takes the FASTQ files, an mRNA reference file, and an AbSeq reference file (if the latter is required) for alignment and annotation. The Whole Transcriptome Analysis (WTA) pipeline takes the FASTQ files, a reference genome, a supplemental reference (if needed), a transcriptome annotation file, and an AbSeq reference file (if the latter is required). See Figure 2.

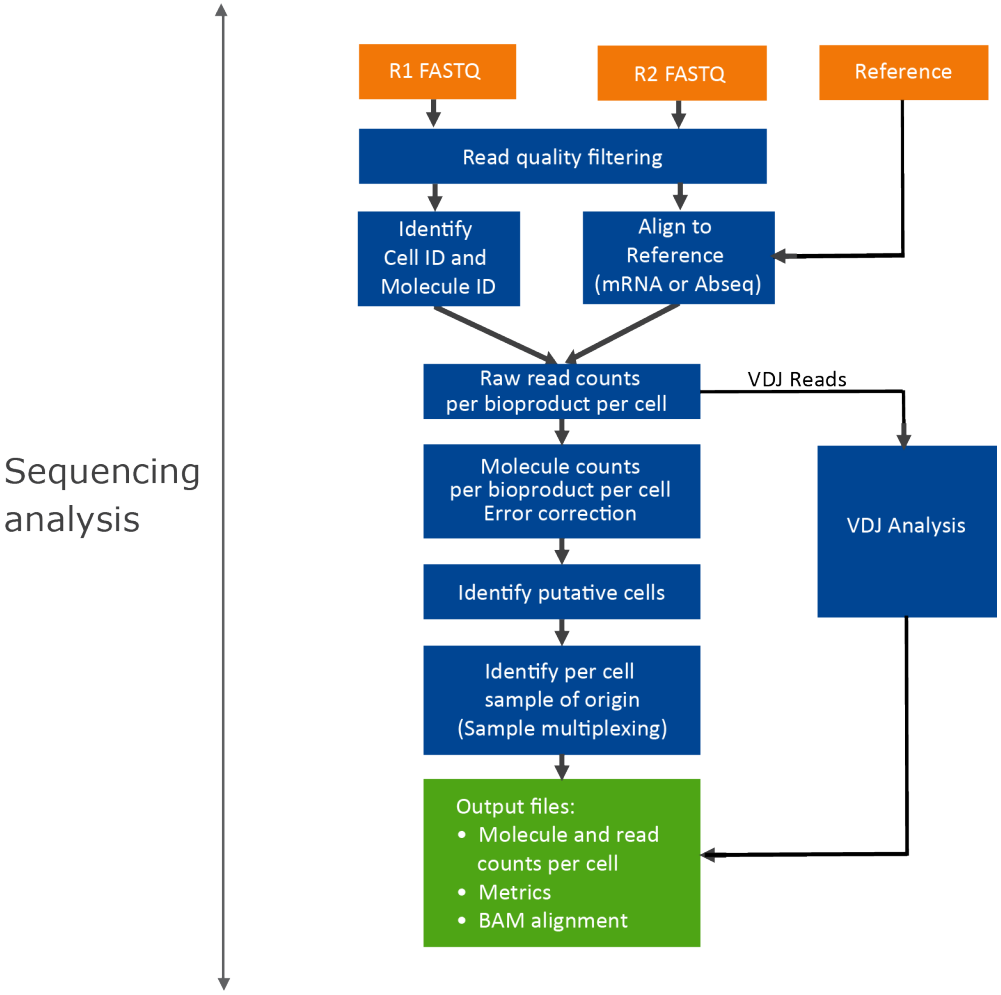


Figure 2 Overview of the steps in the analysis pipeline. For definitions of terms, see Glossary on page 65.

The next sections describe the analysis pipeline step-by-step.

Step 1. Filter by read quality

Filtering criteria

Read 1 artifacts are first removed from read 2. Then, read pairs with low sequencing quality are removed. This step reduces the influence of poor sequencing quality from the metrics that are specific to the BD Rhapsody™ assays.

Read 1 artifacts are removed from read 2 with the following steps:

- Each overlapping read pair is merged by bbmerge with its minentropy setting equal to 18 and minimum overlapping length >= 18.
- The merged read will be split back into a read pair. The merged read will be split according to the bead specific R1 minimum length (described in **Step 2. Annotate R1 reads on page 9**). The bases at the beginning of the merged read up to the R1 minimum length will be assigned to read 1, and the rest will be assigned to read 2.

The following filtering criteria are applied to each read pair after read 1 artifacts are removed:

- Quality filter

Bead	Minimum Read 1 Length	Minimum Read 2 Length	Minimum Mean Quality	R1 Single Nucleotide Frequency	R2 Single Nucleotide Frequency
Original	60	40	20	0.55	0.8
Enhanced 3'	46	40	20	0.55	0.8
Enhanced TCR/BCR	63	40	20	0.55	0.8

- Read length: If the length of the R1 read is less than the bead specific R1 minimum length (described in **Step 2. Annotate R1 reads on page 9**) or the R2 read is <40 bp, the R1/R2 read pair is dropped.
- Mean base quality score of the read: If the mean base quality score of either the R1 read or the R2 read is <20, the read pair is dropped.
- Highest Single Nucleotide Frequency (SNF) observed across the bases of the read: If the SNF is ≥0.55 for the R1 read or the SNF is ≥0.80 for the R2 read, the read pair is dropped. This criterion removes reads with low complexity such as strings of identical bases and tandem repeats.

The thresholds for each filter are determined empirically.

Step 2. Annotate R1 reads

R1 structure

The quality-filtered R1 reads are analyzed to identify the cell label sequences (CLS), common linker sequences (L), Unique Molecular Identifier (UMI) sequence, and if applicable, poly(T) tail. The minimum R1 read lengths for the various beads are:

- Original: 60
- Enhanced 3': 46
- Enhanced TCR/BCR: 63

See Figure 3.

Bead	TCR/BCR Handle	Diversity Insert	CL1	Linker1	CL2	Linker2	CL3	UMI	Capture Sequence
Original	Not Applicable	None	9bp	ACTGGCCTGCGA	9bp	GGTAGCGGTGACA	9bp	NNNNNNNN	18dT
Enhanced 3'	Not Applicable	None, A, GT or TCA	9bp	GTGA	9bp	GACA	9bp	NNNNNNNN	25dT
Enhanced TCR/BCR	ACAGGAAACTCATGGTGCGT	None	9bp	AATG	9bp	CCAC	9bp	NNNNNNNN	TATGCGTAGTAGGTATG

Figure 3 Structure of R1 read

Cell label

Information of the cell label is captured by bases in three sections (CLS1, CLS2, CLS3) along each R1 read. Two common sequences (L1, L2) separate the three CLSs, and the presence of L1 and L2 relates to the way the capture oligonucleotide probes on the beads are constructed. By design, each CLS has one of either 96 or 384 predefined sequences (depending on bead version), which has a Hamming distance of at least four bases and an edit distance of at least two bases apart. A cell label is defined by the unique combination of predefined sequences in the three CLSs. Thus, the maximum possible number of cell labels is either 96^3 or 384^3. In the final data tables, the three part cell label is converted to a single integer index between 1-384^3

Reads are first checked for perfect matches in all three pre-designed CLS sequences at the expected locations, and reads with perfect matches are kept.

The remaining reads are subjected to another round of filtering to recover reads with base substitutions, insertions, and deletions caused by sequencing errors, PCR errors, or errors in oligonucleotide synthesis.

UMI

By design, the UMI is a string of eight randomers immediately downstream of CLS3. For reads with insertions or deletions within the CLSs, the UMI sequence is eight bases immediately following the end of the identified CLS3.

Step 3. Annotate R2 reads

Criteria for a valid R2 read

Targeted assays:

For targeted assays, the pipeline uses Bowtie2 to map the filtered R2 reads to the reference panel sequences. Option `--norc` is enabled to map all of the reads only to the forward strand of the provided reference.

Targeted assays:

For targeted assays, an R2 read is a valid alignment if all of these criteria are met:

- The R2 alignment begins within the first five nucleotides for mRNA, first 15 nucleotides for AbSeq, and first 25 nucleotides for Sample Tags. This criterion ensures that the R2 read originates from an actual PCR priming event.
- The length of the alignment match (can be a match or mismatch) in the CIGAR string is >=37 for mRNA >=25 for AbSeq and >=40 for Sample Tags. A CIGAR (Compact Idiosyncratic Gapped Alignment Report) string is a sequence of base lengths to indicate base alignments, insertions, and deletions with respect to the reference sequence.
- The read does not align to phiX174.

WTA assays:

For WTA assays, the pipeline uses STAR to map the filtered R2 reads to the transcriptome. By default, alignments to both exons and introns are used. Including reads that align to introns may increase sensitivity, resulting in an increase in molecule counts and the number of genes per cell for both cellular and nuclei samples. Reads that align to introns may indicate the presence of unspliced mRNAs and are also useful in the study of nuclei and RNA velocity.

An R2 is a valid gene alignment if all of these criteria are met:

- The sum of the CIGAR alignment matches must be >= 25.
- The read aligns uniquely to an exon or intron of a bioproduct in the reference.
- The read does not align to phiX174.
- If "Exclude Intronic Reads" option is selected, read must align to exon.

Step 4. Combine information from R1 and R2 annotations

Retain R1 and R2 reads

Read pairs with a valid R1 read and a valid R2 read are retained for further analysis. A valid R1 read requires identified CLSs, and a UMI sequence with non-N bases.

A valid R2 read must uniquely map to a bioproduct in the reference. For targeted, it must also have the correct PCR2 primer sequence at the start and an alignment match sufficient in length.

Step 5. Annotate molecules

Collapse reads into raw molecules

Reads with the same cell label, same UMI sequence, and same bioproduct are collapsed into a single raw molecule. The number of reads associated with each raw molecule is reported as the *raw adjusted sequencing depth*.

Remove artifact molecules using RSEC and DBEC UMI adjustment algorithms

PCR and sequencing often generate errors. If the error occurs within the UMI sequence, the R1/R2 read pair is called a unique molecule but is, in fact, an artifact. Artifact molecules contribute to an over-estimated molecule count of a gene in a cell. As sequencing depth increases, the number of raw molecules rises and never plateaus due to these artificial molecules.

To remove the effect of UMI errors on molecule counting, BD Biosciences has developed a set of UMI adjustment algorithms. UMI errors that are single base substitution errors are identified and adjusted to the parent UMI barcode using recursive substitution error correction (RSEC). For targeted sequencing analysis, other UMI errors derived from library preparation steps or sequencing base deletions are later adjusted using distribution-based error correction (DBEC).

Note that targeted sequencing analysis uses RSEC and DBEC, while WTA sequencing analysis uses RSEC only for mRNA libraries and RSEC and DBEC for AbSeq libraries.

Figure 4 shows the targeted workflow using both the RSEC and DBEC algorithms on data generated from BD Rhapsody™ targeted assays. Figure 5 shows the WTA workflow using RSEC on the mRNA libraries and both RSEC and DBEC on the AbSeq libraries. Figure 6 shows an example on how the RSEC and DBEC algorithms are applied to correct the apparent counts of molecules.

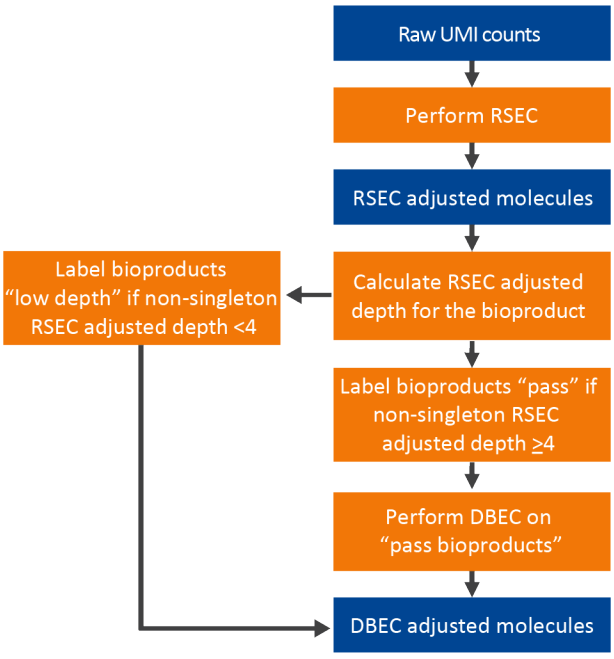


Figure 4 Workflow of UMI count adjustment for targeted assays

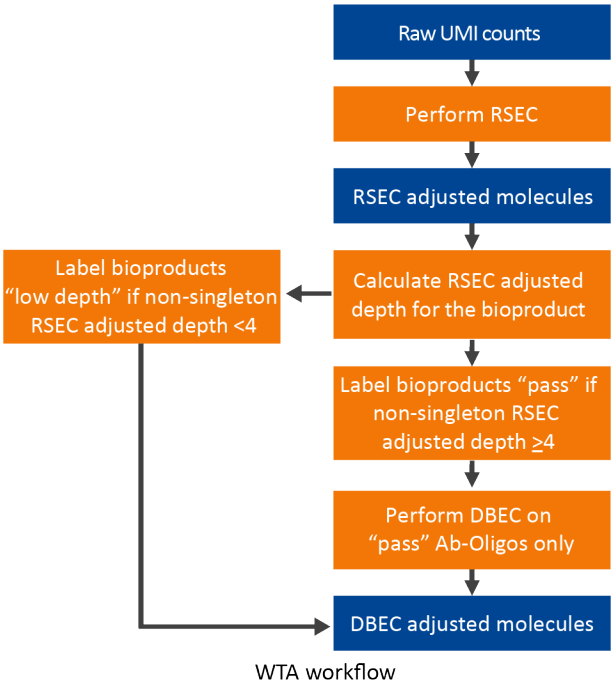


Figure 5 Workflow of UMI count adjustment for WTA assays

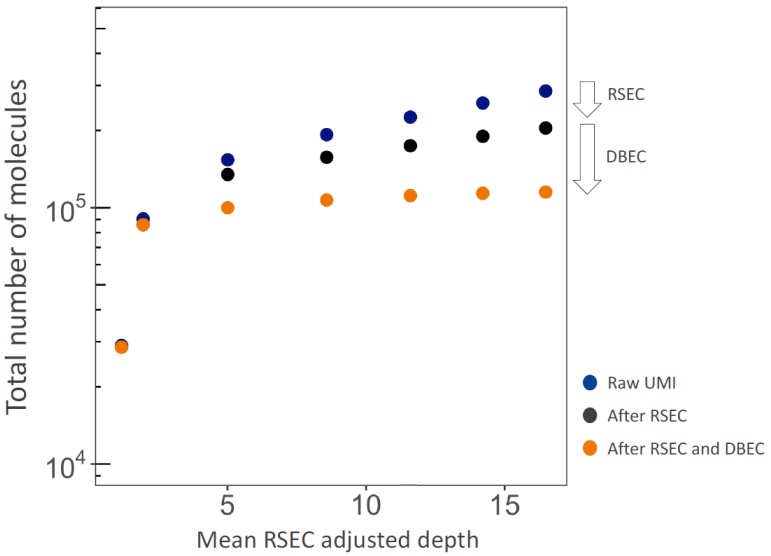


Figure 6 Applying RSEC and DBEC to an example dataset. For targeted sequencing analysis, if we consider only raw UMIs, the apparent total number of molecules continues to rise with sequencing depth, because the presence of sequencing and PCR errors contribute to unique UMIs. RSEC removes artifact molecules from single base substitutions in the UMI sequence. Further adjustment by DBEC removes artifact molecules originating from PCR errors. As a result, the number of molecules stabilizes with additional sequencing, indicating the library is sequenced to saturation.

Collapse molecules that differ by one base in the UMI sequence using RSEC

RSEC considers two factors in error correction: 1) similarity in UMI sequence and 2) raw UMI coverage or depth. See Figure 7.

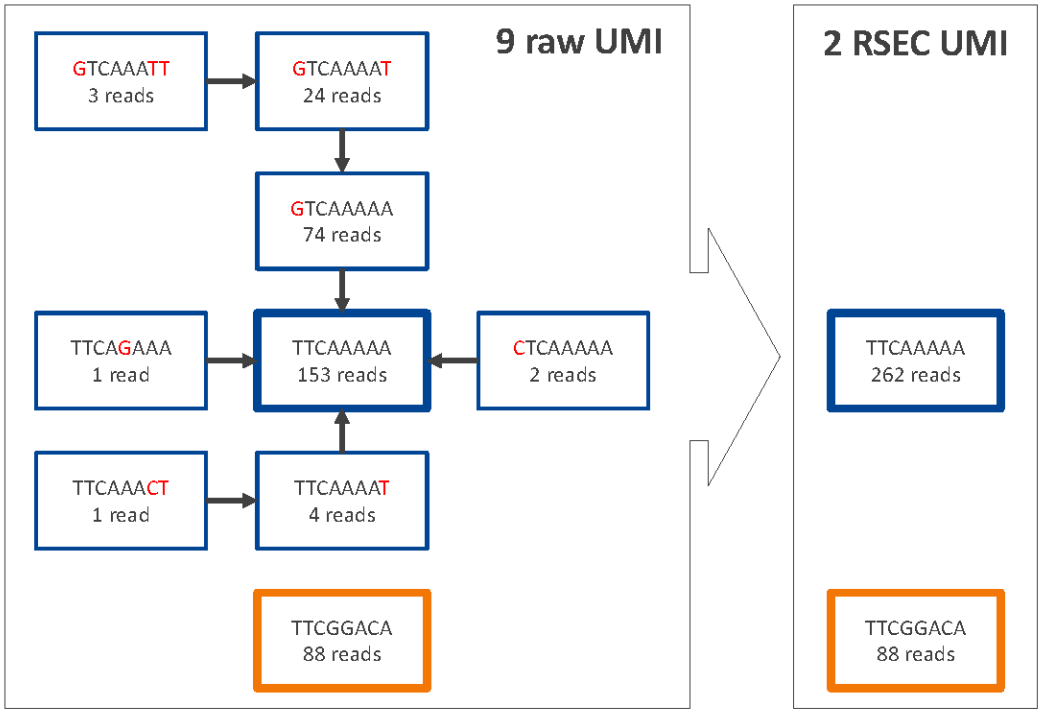


Figure 7 Example of the RSEC algorithm. Nine raw UMIs are collapsed into two UMIs.

For the molecules from each combination of cell label and bioproduct, UMIs are connected when their UMI sequences are matched to within one base (Hamming distance = 1). For each connection between UMI x and y, if Coverage(y)> 2 * Coverage(x) – 1, then y is the Parent UMI and x is the Child UMI. Based on this assignment, child UMIs are collapsed to their parent UMI. This process is recursive until there are no more identifiable parent-child UMIs for the bioproduct. See Figure 7.

The number of reads for each child UMI is added to the parent, so no reads are lost. The sum of the reads is the *RSEC-adjusted depth* of the *RSEC-adjusted molecule*.

Adjust molecule counts by DBEC

The RSEC-adjusted molecule counts are further corrected by DBEC, depending on assay type. For Targeted assays, DBEC is applied on all bioproduct types (mRNA and AbSeq). For WTA assays, DBEC is applied only to AbSeq targets.

DBEC is applied on a per-bioproduct basis. The algorithm is based on the assumption that the pre-amplified set of molecules of the same bioproduct, regardless of the cell of origin, is subject to the same amplification efficiency and, therefore, should have similar read depth. Artifact molecules created later in the PCR cycles, such as those derived from PCR chimera formation, will likely have less read depth.

DBEC considers the distribution of RSEC-adjusted depth distribution, not UMI sequence. The sequencing depth of RSEC-adjusted molecules for each bioproduct is a bimodal distribution. See Figure 8. The lower mode of the distribution likely represents artifact molecules, and the upper mode likely represents true molecules. The

algorithm fits two negative binomial distributions to statistically distinguish between the two modes. Molecules in the upper mode are retained (*DBEC-adjusted molecules*), while the molecules in the lower mode are discarded. The average depth of the molecules in the upper mode is known as the *DBEC-adjusted depth*, and the depth of molecules in the lower mode is the metric *error depth*. The cutoff between the two modes is the *DBEC minimum depth*.

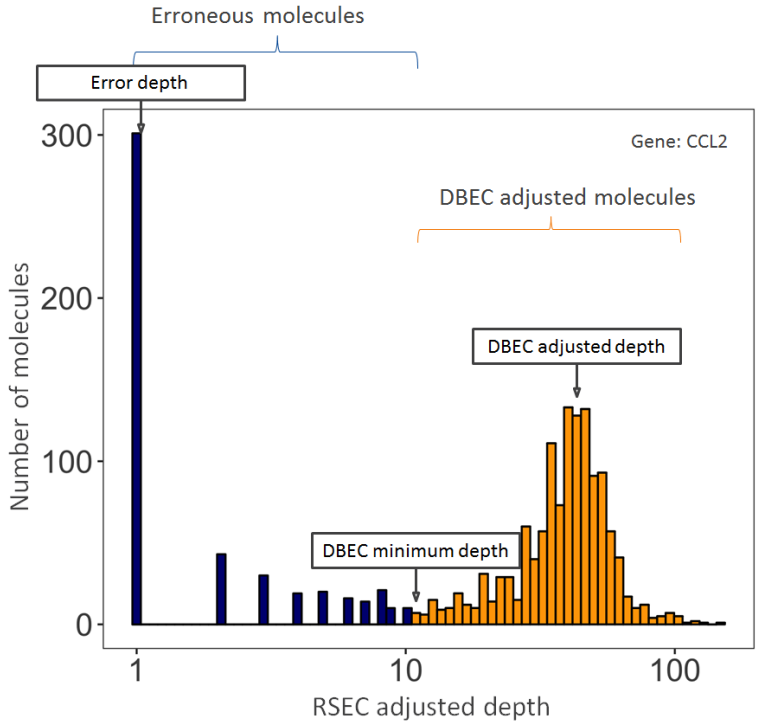


Figure 8 Example of the DBEC algorithm for gene CCL2. Counts under the orange bars are kept and labeled as DBEC-adjusted molecules. Counts under the blue bars are labeled as erroneous molecules and are discarded. The error depth and DBEC-adjusted depth arrows point to the respective average depths.

DBEC is applied to bioproducts with an average non-singleton RSEC sequencing depth ≥4. This means that the depth is calculated after removing RSEC UMIs with only one representative read. According to the Poisson distribution, if the average UMI depth is <4, more signal UMIs are removed than error UMIs. As a result, a bioproduct is marked as *pass* if its average RSEC depth ≥4 and is subject to DBEC. Otherwise, it is marked *low depth* and bypasses DBEC. If no count is associated with the bioproduct, it is labeled as *not detected*.

DBEC removes molecules and the reads associated with the removed molecules from consideration in downstream analyses. The percentage of reads retained by DBEC is reported together with the other pipeline metrics.

The RSEC and DBEC metrics associated with each bioproduct are reported in the file, <sample_name>_ Bioproduct_Stats.csv.

Step 6. Determine putative cells

Excessive cell labels

In theory, the number of unique cell labels detected by the bioinformatics pipeline should be similar to the number of cells captured and amplified by the BD Rhapsody™ workflow. However, various processes throughout the workflow can introduce noise that contribute to excessive cell labels generated during sequencing analysis, including:

- Hybridizing polyadenylated [poly(A)] oligonucleotides to beads residing in neighboring wells when the cell lysis step is too long
- Underloading beads in BD Rhapsody™ Cartridges resulting in cells without beads and the RNA from the cells diffusing to adjacent wells
- Experiencing low-level contamination during oligonucleotide and bead synthesis
- Generating errors during the PCR amplification steps of the workflow

To distinguish cell labels associated with putative cells from those associated with noise, a multi-step algorithm was designed for filtering cell labels. See Figure 9.

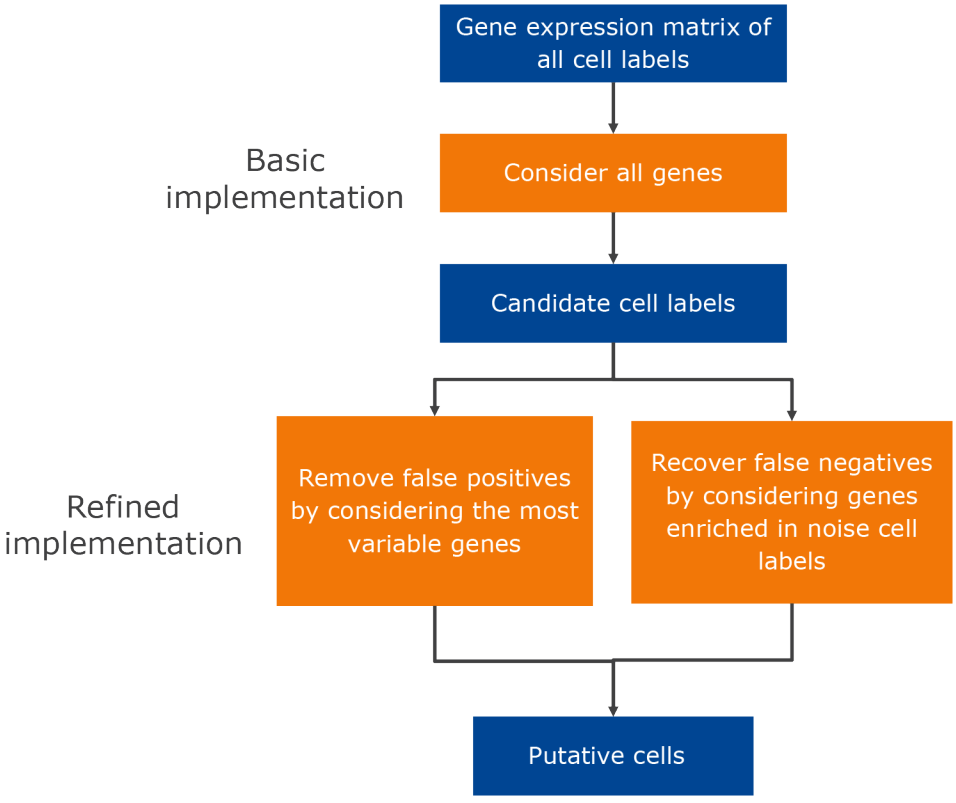


Figure 9 Workflow for determining putative cells.

Putative cell identification using second derivative analysis (basic implementation)

The principle of the cell label filtering algorithm is that cell labels from actual cell capture events should have many more reads associated with them than noise cell labels. By default, read counts from mRNA molecules are used to determine the number of putative cells. Putative cell calling can also be performed using the basic implementation on the AbSeq read counts. All reads associated with all DBEC-adjusted molecules (RSEC-adjusted molecules for molecules that did not undergo DBEC correction) from the selected bioproduct type (mRNA by default) are taken into account. The number of reads (post-DBEC) of each cell is plotted on a log10-transformed cumulative curve, with cells sorted by the number of reads in descending order. See Figure 10, left. In a typical experiment, a distinct inflection point is observed, indicated by the red vertical line. The algorithm finds the minimum second derivative along the cumulative read curve as the inflection point. See Figure 10, right. Cell labels to the left of the red vertical line (Figure 10, left) are most likely derived from a cell capture event and are considered as signal (labeled as cell labels set A or candidate cell labels). The remaining cell labels to the right of the red line (Figure 10, left) are noise. Up to this point, the analysis is the basic implementation of the second derivative analysis.

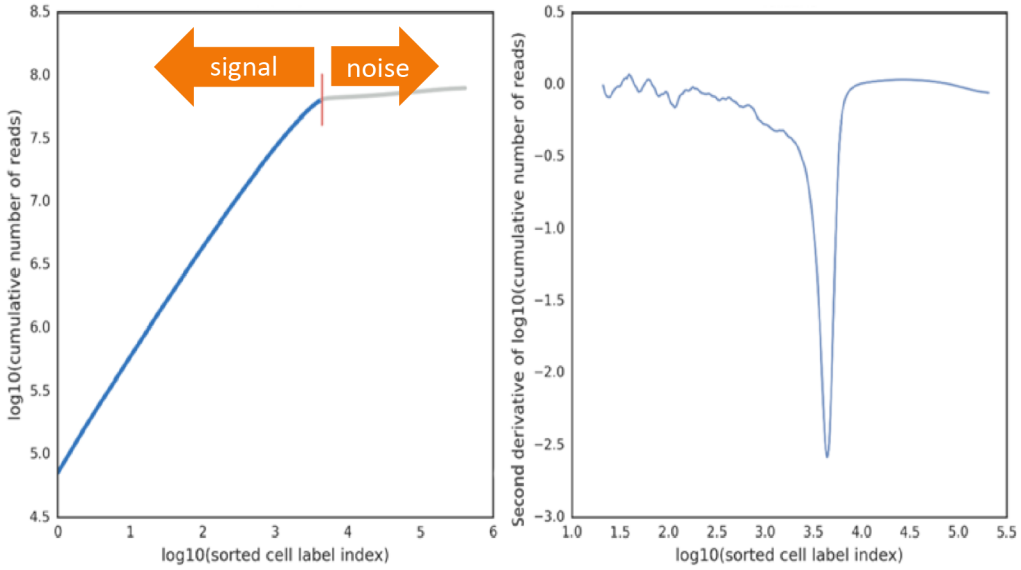


Figure 10 Results of the basic implementation of the second derivative analysis applied to a typical BD Rhapsody™ library.

If every cell in the sample is well represented by molecules from library preparation, there is only one inflection point. The number of reads of the putative cells is a single distribution well separated from the noise distribution.

There are situations, however, when a sample contains cells with a very wide range of number of molecules. If sub-populations of cells with high and low mRNA content are considerably large, multiple inflection points can be observed. Example scenarios include biological samples such as peripheral blood mononuclear cells (PBMCs) with plasma cells being much larger and active carrying thousands of molecules compared to lymphocytes being smaller and less active carrying tens of molecules (see Figure 11A), or artificial mixtures of cell line cells and primary cells (see Figure 11B). The basic implementation of the second derivative analysis chooses the inflection point that includes all distributions beyond the usual noise distribution. Specifically, inflection points are considered valid if the second derivative minimum corresponding to the inflection point is at least half as deep as the global minimum and is ≤–0.3. The smoothing window of the second derivative curve increases until

there are two valid inflection points. The inflection point corresponding to the larger cell number is deemed the better one.

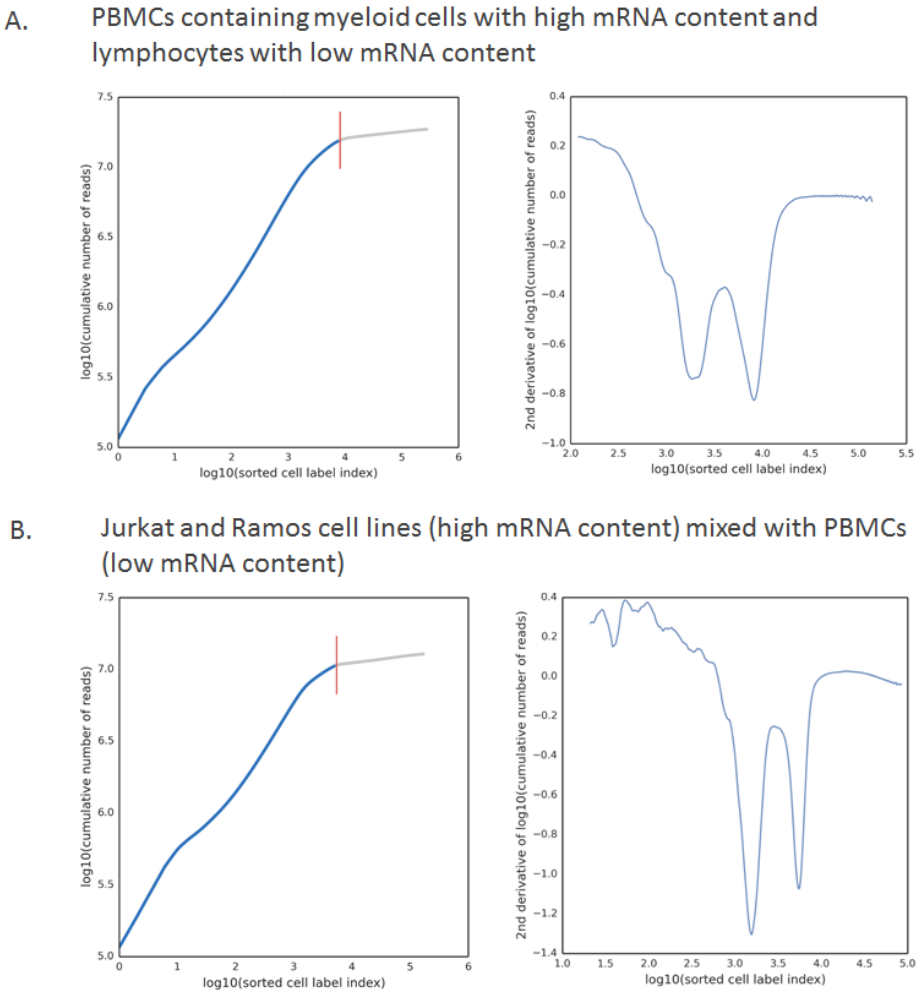


Figure 11 Results of the basic implementation of the second derivative analysis on libraries with very different levels of mRNA content. A. PBMCs with myeloid (high mRNA content) and lymphoid (low mRNA content) cells. B. Mixture of Jurkat and Ramos cells (cell lines, high mRNA content) and PBMCs (low mRNA content). Both libraries were analyzed with the BD Rhapsody™ Immune Response Panel Hs (human).

Removing false positives and recovering false negatives (refined implementation)

In some cases, the basic implementation of the second derivative analysis might include small numbers of false positive and false negative cell labels. Additional refinement steps are implemented to identify these false positive and false negative cell labels in order to generate a final set of cell labels for further analysis. This refined implementation is applied only when mRNA read counts are used for putative cell calling.

Removing false positives

Consider the case where the chosen inflection point includes the populations of cell labels with wide ranges of number of reads per cell label. Then, the signal population with lower reads per cell label might also include

noise cell labels derived from residual mRNA molecules from the cells with very high mRNA content. The number of reads associated with these noise cell labels derived from high-expressing cells can be indistinguishable from low-expressing cells, which have similar reads per cell.

Since these false positive cells can be hard to identify with reads alone, the relative expression profile of cell labels can be used to identify them. For example, a false positive cell label that is derived from a high mRNA-expressing, true positive cell label would likely have a similar expression profile but with a lower read signal. Therefore, a second derivative analysis is done on the most variable genes to identify these false positive cell labels.

The most variable gene expression is defined by a process similar to that described by Macosko, EZ, et al. [see References on page 63]:

a. Log-transform read counts of each gene within each cell to get the gene expression: $\log_{10}(\text{count} + 1)$.
b. Calculate the mean expression and dispersion (defined as variance/mean) for each gene.
c. Place genes into 20 bins based on their average expression.
d. Within each bin, calculate the mean and standard deviation of the dispersion measure of all genes, and then calculate the normalized dispersion measure of each gene using the following equation:

   Normalized dispersion = (dispersion – mean)/(standard deviation)

e. Apply a cutoff value for the normalized dispersion to identify genes for which expression values are highly variable even when compared to genes with similar average expression.

A second derivative analysis is applied on variable gene sets defined by a different cutoff value for the normalized dispersion to derive the cell label filtered set B. For each dispersion cutoff, the noise cell labels are determined as A – B. For instance, for three cutoff values, noise cell labels are N1 = A – B1, N2 = A – B2, and N3 = A – B3, where the minus sign represents the set difference. The common noise cell labels detected among N1, N2, and N3 are subtracted from cell labels set A. The resultant set is denoted as cell label filtered set C = A – intersection(N1, N2, N3).

Recovering false negatives

Cells with low numbers of molecules might be missed by the basic implementation of the second derivative analysis algorithm, because a cell subset might express very few of the genes in the gene list. The cell labels carry a very low number of reads, and the size of the cell population is small enough that their cell labels do not form a distinct second inflection point. These cell labels might be mistaken as noise.

If there are genes specific to the false negative cell label subset (for example, marker genes), they can be identified by comparing the number of reads for each gene from all detected cell labels to those from cell labels deemed as signal. The assumption is that the relative abundance of reads for each gene from all of the noise cell labels should be no different than that from all of the cell labels considered as signal. If a specific cell subset is missed initially, there is a set of genes that appears as enriched in the noise cell labels in the basic implementation.

This enriched set of genes is detected by the following steps:

a. For each gene, calculate the total read counts from all detected cell labels and from cell labels in set C.
b. Identify the genes that have the biggest discrepancy in representation by cell labels in set C versus all cell labels. This is done by plotting and finding the line of best fit to detect the genes with the largest residuals at least one standard deviation away from the median of residuals of all genes. See Figure 12.

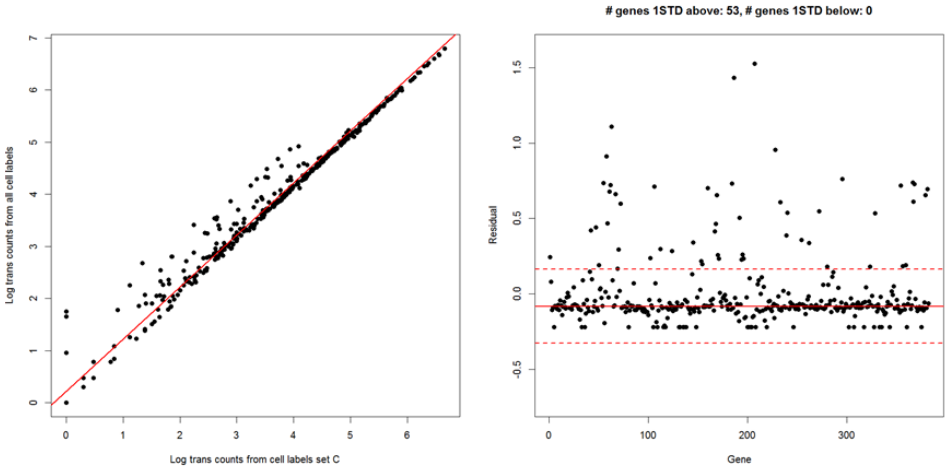


Figure 12 A. and B. Detecting genes enriched in noise as determined by the basic implementation of the second derivative analysis. Each dot represents a gene. B. The two red dashed lines correspond to one standard deviation above and below the median (red solid line). In this example, 53 genes are enriched in the noise population.

The second derivative analysis algorithm is run again with this enriched set of genes. The recovered cell labels (*cell label filtered set D*) are combined with cell labels in set C to form set E. As a final cleanup step, cell labels carrying less than the minimum threshold number of molecules are removed. The number of cell labels in the final set is *the number of putative cells*.

Identify Protein Aggregates from AbSeq Read Counts

When identifying putative cells using the AbSeq read counts, the basic implementation may include a small number of false positive cell labels due to protein aggregates. Putative cells identified with high expression across most AbSeq targets are considered protein aggregates. The protein aggregate status for each putative cell can be found in the <sample_name>_Protein_Aggregates_Experimental.csv file. The cell label is marked True if it is considered a protein aggregate and False if not.

Reporting putative cells

The category of each cell label is listed in the file <sample_name>_Putative_Cells_Origin.csv. The cell label is marked *basic* if it is considered a putative cell in the basic implementation when the second derivative analysis is run using data from all bioproducts in the bioproduct list. A cell label is marked as *refined* if it is considered a putative cell in the refined implementation and is a recovered false negative. In most cases, most putative cell labels originate from the basic implementation. See Putative cells origin on page 48.

Step 7. Determine the sample of origin (sample multiplexing only)

Sample multiplexing option

Multiple samples of cell suspension can be loaded into a BD Rhapsody™ Cartridge using a BD® Single-Cell Multiplexing Kit. Each sample is labeled with a separate Sample Tag from the kit. The human and mouse sample kits provide up to 12 species-specific sample tags. The flex sample kit provides up to 24 species and cell type agnostic sample tags.

When you start the BD Rhapsody™ Analysis pipeline, you can select the sample multiplex option. You can associate a name with a Sample Tag before the pipeline starts, and the specified sample names will be used in the output files.

To account for every Sample Tag, each Sample Tag sequence in the kit is considered during pipeline analysis, whether the Sample Tags are used in the experiment or specified with a sample name.

The pipeline automatically adds the Sample Tag sequences to the FASTA reference file. Reads that align to a Sample Tag sequence and associate with a putative cell are used to identify the sample for that cell.

Sample determination algorithm

The algorithm first identifies high quality singlets. A high quality singlet is a putative cell where more than 75% of Sample Tag reads are from a single tag. When a singlet is identified, the counts for all the other tags are considered Sample Tag noise. See Figure 13. Sources of low-level noise can be PCR and sequencing errors and residual Sample Tag labeling during cell preparation.

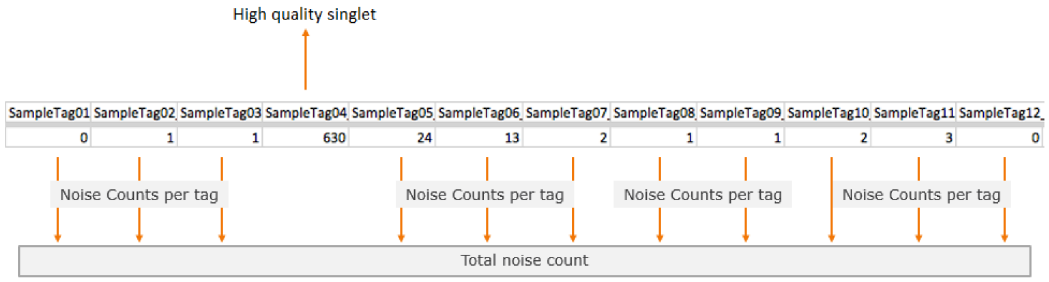


Figure 13 Example of Sample Tag read counts for a putative cell that is considered a high quality singlet, labeled SampleTag04. All of the other Sample Tag counts are recorded as separate noise counts and are summed to find the noise read count for that putative cell.

The minimum Sample Tag read count for a putative cell to be positively identified with a Sample Tag is defined as the lowest read count of a high quality singlet for that Sample Tag. See Figure 14.

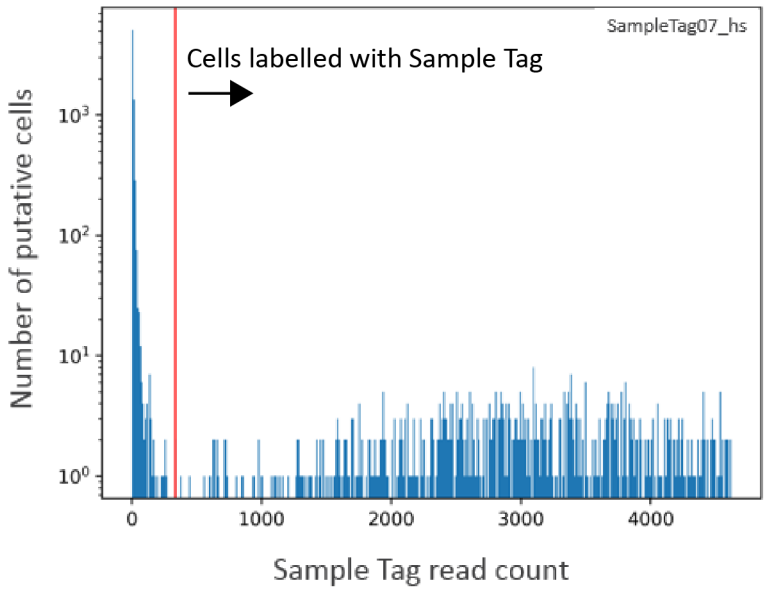


Figure 14 Histogram of number of Sample Tag read count per putative cell for one of the Sample Tags. The red vertical line indicates the threshold of minimum Sample Tag read count. Putative cells with Sample Tag read counts greater than the threshold (to the right of the red line) are considered labelled with this Sample Tag. In addition to singlets, these putative cells can include multiplets, which are cell labels associated with more than one Sample Tag.

The percentage of noise contribution of each Sample Tag for all cells is calculated by dividing the total per tag noise by the total overall noise. In addition, the total amount of noise versus the total Sample Tag count per putative cell is recorded so that a trend line can be established to estimate the total per-cell noise given an observed number of total Sample Tag count for a cell. See Figure 15. The level of antigen expression across cells can vary, contributing to variation in Sample Tag count per cell. Generally, cells with higher total Sample Tag counts have higher noise Sample Tag counts.

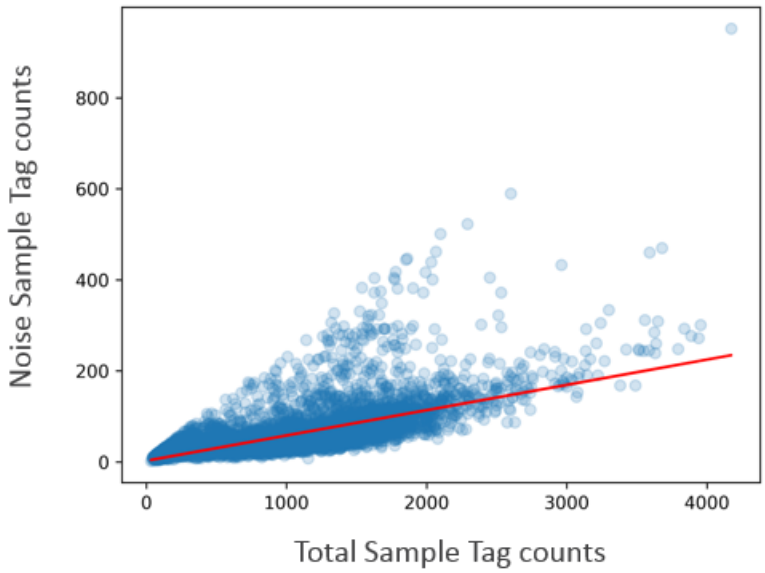


Figure 15 Overall noise profile where each dot is a cell. A trend line (in red) is fitted and used to establish the expected amount of noise given a total Sample Tag count. Cells that are off the trend line are likely multiplets.

To improve sample determination and recover singlets that are not initially considered high quality, the algorithm subtracts the expected number of per-cell noise counts from each Sample Tag. The total expected per-cell noise, derived from the trend line, is multiplied by the percentage of noise contribution of each Sample Tag to determine the expected noise per Sample Tag.

After subtracting the expected per tag noise, any Sample Tag that has a count higher than its minimum read count is called for that cell, and the putative cell is considered a *called* cell.

When the counts of two or more Sample Tags exceed their minimum thresholds, then that putative cell is called as a cross-sample *Multiplet*, indicating more than one actual cell in the microwell, and the cells are of different samples of origin. Some putative cells might not have enough Sample Tag counts to definitively call their sample of origin, and those are labeled as *Undetermined*.

Reporting sample origin

If you chose the sample multiplexing option, the main top-level RSEC and DBEC data tables contain counts for putative cells from all samples combined. The sample of origin for each putative cell is listed in the file <sample_name>_Sample_Tag_Calls.csv. This file can be used to annotate the combined data tables. The file <sample_name>_Sample_Tag_Metrics.csv reports the metrics from the sample determination algorithm. Data tables and metric summary for each sample are output in folders contained in <sample_name>_Sample_Tag<number>.zip.

Step 8. Generate expression matrices

Reporting RSEC and DBEC metrics

RSEC-adjusted molecule counts and associated reads of each bioproduct for each putative cell and DBEC-adjusted molecule counts and associated reads are presented in either .csv or .st format. See Expression data on page 47 and Data tables on page 46.

Step 9. Annotate BAM

Annotating BAM

The BAM file output by Bowtie2 or STAR is further annotated to summarize the results of the BD Rhapsody™ Analysis pipeline. The table lists the tags appended to the annotation of each read. For BAM tags, see BAM and BAM Index on page 45, samtools.github.io/hts-specs/SAMv1.pdf, bowtie-bio.sourceforge.net/bowtie2/manual.shtml#sam-output, and github.com/alexdobin/STAR/blob/2.5.2b/doc/STARmanual.pdf.

Step 10. TCR and BCR analysis (if applicable)

TCR and BCR overview

In combination with other BD Rhapsody™ assays, optional protocols and products enable the generation of sequencing libraries specific to T- and B-Cell Receptors (TCR and BCR). When enabled, the BD Rhapsody™ Analysis pipeline can use the reads from these libraries to assemble contigs corresponding to rearranged TCR and BCR chain mRNA. These contigs are then analyzed to identify single-cell level VDJ gene segments, complementary determining regions (CDRs), read and molecule counts, and per cell type chain-pairing. TCR and BCR analysis is supported for human and mouse.

Major TCR and BCR pipeline steps

- Identify reads derived from TCR or BCR mRNA
- Assemble reads into contigs
- Annotate contigs with VDJ gene segment information
- Select dominant contigs, chain family, cell type
- Error correction and contig trimming
- TCR and BCR output files on page 53

Identify reads derived from TCR or BCR mRNA

As described in Step 3. Annotate R2 reads on page 10, reads are aligned against a reference sequence to determine their biotype and identity (for example, which gene, AbSeq, sample tag, or VDJ gene segment).

For the WTA assay, in combination with TCR and BCR, the pipeline will identify TCR or BCR reads which align to known VDJ gene egments in the transcriptome, with the appropriate orientation. Known VDJ segments are those with a transcriptome GFF “gene_biotype” starting with “TR_” or “IG_”.

For the targeted assay in combination with TCR and BCR, the pipeline automatically adds species appropriate TCR and BCR gene segments to the FASTA reference file. These gene segments are derived from the same Gencode transcriptome GFF as is used in the WTA assay.

Reads that align to TCR or BCR gene segments are grouped and separated from the reads aligning to other biotypes. These reads then only go through the procedures described in the remainder of this section, and not the steps described previously for other biotypes.

Assemble reads into contigs

To generate the full variable region of a TCR or BCR sequence, or the consensus CDR3 sequence, short reads must be assembled. Read assembly operates by looking for similarities and overlaps between reads that suggest they originate from the same original sequence. Aligning and stitching these reads together can allow for the creation of longer contigs from short reads, and correct randomly distributed sequencing errors.

Reads identified as TCR or BCR derived, are prepared for assembly with trimming and UMI error correction. First, the 3’ end of reads are trimmed with a quality score threshold of 20. It’s important that the reads going into assembly be of high quality, so that reads can be correctly aligned, and a valid consensus sequence can be generated. Next, reads are also trimmed based on bead capture sequences to remove artifacts from the BD Rhapsody™ cell label sequence. These sequences could interfere with the correct assembly, and may be

found at the 3' end of TCR or BCR reads if they were derived from short amplicons. Then, reads undergo UMI error correction, grouped by their cell ID and the TCR or BCR chain type determined by initial alignment (for example, TCR-Alpha, IG-Kappa...). This UMI error correction step uses the same RSEC algorithm described previously.

To begin assembly, reads are grouped by their cell ID and chain type. These read groups are sent through a software package for transcript assembly called Trinity, which generates a list of contig sequences. Then, the reads are aligned back to the newly generated contigs, in order to produce read and molecule counts for each contig. Multiple contigs from each cell represent the rearranged VDJ mRNA sequences, for instance, of TCR Alpha and TCR Beta chains.

Annotate contigs with VDJ gene segment information

All contigs generated by the assembly step are analyzed to identify V, D, J, and C gene segments, complementarity determining regions (CDR1-3), framework regions (FR1-4), productivity (lack of stop codons), if contig is full length, and protein sequence. This analysis is accomplished with a software package called IGBlast and with alignments using Bowtie2.

A contig is removed from further analysis if a V or J gene selection is of low quality, indicated by an e-value score greater than 10^(-3) (lower is better). A contig is considered "full length" when there is amino acid sequence defined for each framework (1-4) and CDR (1-3) region. For the "full length" metric, FR1 and FR4 may be partial, but the overall contig is still considered full length.

Select dominant contigs, chain family, cell type

For each cell and chain type, a dominant contig is selected to facilitate reporting, metrics, and downstream analysis. The selection of a dominant contig follows these rules:

- Contigs containing a CDR3 are considered in the top tier.
- To break any ties, the contigs are then sorted in order of: highest molecule count, highest read count, best V-segment e-value quality score, and productivity.
- All contigs are also output in separate files, and are still available to be analyzed.

Secondary contigs can be generated due to biological reasons, like dual expression of alpha or beta TCR chains expression or assay-based reasons, like: sequencing errors, transcription errors, contaminating reads from other mRNAs, cell multiplets, and mis-assembly.

For cells expressing both TCR alpha/beta and TCR gamma/delta, a single chain family is selected for the final output file, but all data is still available in the uncorrected output. To select the chain family, expression of both alpha and beta or gamma and delta is preferred. Then, if all 4 chains or one of each chain has expression, the family with the highest combined molecule count is selected (alpha+beta vs gamma+delta).

During any BD Rhapsody™ assay in combination with TCR and BCR, putative cell determination is still based on 3' gene expression from targeted or WTA data. This is more accurate than creating separate putative cell identifications for the TCR and/or BCR libraries. The VDJ metrics contain a breakdown of metrics by cell type. Cell types are determined in one of two ways. The pipeline contains an experimental immune cell type classifier that uses a machine learning model developed on human PBMCs. This method will only work when the targeted or WTA gene expression datatables contain counts for a set of 100 core genes relevant to the model.

The TCR and BCR algorithms contain a simple fallback for cell type determination in the case of human data where gene expression was not available for those 100 genes, or for mouse data:

- A putative cell with 2x more TCR molecules than BCR molecules, (or only TCR data) is a T cell.
- A putative cell with 2x more BCR molecules than TCR molecules, (or only BCR data) is a B cell.
- A putative cell without a 2x difference, or one without any TCR or BCR data is unknown.

Cell type determination by the fallback mechanism may be different in unfiltered data vs corrected data.

Contig trimming and error correction

Dominant contigs from putative cells undergo two additional steps before final reporting. First, the 3' end of contigs are trimmed based on the identified constant region and the known primer sequence it contains. Any assembled sequence 3' of the primer sequence, within the constant region, is not consequential to the VDJ region, and likely assembled in error.

To improve specificity, dominant contigs from putative cells undergo a final round of error correction on a per chain type basis: for each of TCR alpha, beta, gamma, delta, and BCR heavy, kappa, and lambda. This distribution style error correction assumes that each individual chain type from T or B cells will amplify at similar rates, and thus would end up with similar numbers of reads per cell-chain for a real TCR or BCR expressing cell. Artifact molecules created with non-T or non-B cell labels in late PCR cycles, such as those derived from PCR chimera formation, will likely have fewer reads. The algorithm is multimodal aware, so that if there are 2 positive populations for a particular chain, they should both be kept (for example, Naïve B cells and plasma B cells with different reads per cell in the same experiment).

First, there is a check to determine if each chain type has a read depth of at least 4 reads per cell. If, not, then error correction does not proceed for that chain type. Next, a histogram of the reads per cell from each chain type is generated, and a multimodal distribution is modeled on each. A threshold is set at the local minima between the first and second modes, and on a per chain basis, any TCR or BCR data from cells whose reads counts are in the lowest mode are removed.

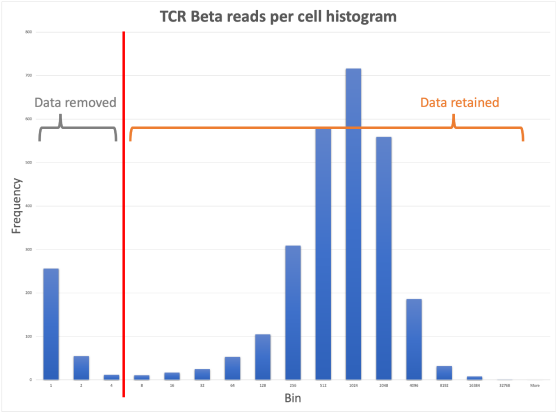


Figure 16 Contig error correction is based on modeling a multimodal distribution on the number of reads per cell for each chain type.

Untrimmed contigs and contigs before error correction are still available in an unfiltered contigs output file.

Step 11. Generate summary

Metrics summary

A summary .csv file documenting the metrics of each of the analysis steps is generated. See Metrics summary on page 37.

Pipeline report HTML

A pipeline report HTML file is generated and contains the results from the sequencing analysis pipeline. The pipeline report is a stand-alone HTML file that requires no internet connection making it easy to share with collaborators. The pipeline report contains several graphs to help visualize the results. The metrics that are shown in the report are the same as those found in the Metrics Summary CSV file. There are also helpful tooltips for each metric that describe the specific metric in more detail. The pipeline report also contains the pipeline inputs that were specified for the sequencing analysis, allowing you to re-run the analysis using the same inputs.

Summary section

The pipeline report starts out with the summary information at the top with the most important metric results. The number of putative cells is shown in the center. On the left side of the summary, some key library specific metrics such as the number of reads in the FASTQ, the percentage of reads that passed all the quality filters, and the percentage of reads that aligned uniquely are shown. On the right side of the summary, some key bioproduct type metrics such as how many reads were aligned, the mean reads per cell, and the mean molecules per cell are highlighted.

Graph section

The graph section has several interactive graphs highlighting some of the most important results from the analysis.

Single Bioproduct Expression

The Single Bioproduct Expression graph displays a tSNE on the left and a histogram on the right for individual bioproducts. Each dot on the tSNE represents a putative cell and is colored by the log 10 expression of the selected AbSeq target or mRNA gene. The histogram shows the distribution of expression for all cells for the selected AbSeq target or mRNA gene. By default, the bioproduct with the highest expression is selected in the dropdown list. The AbSeq targets and mRNA genes are sorted by total expression (highest to lowest) separately. The sorted AbSeq targets are shown first in the dropdown list followed by the sorted mRNA genes. For larger experiments, only the most highly, widely, and variably expressed genes plus all AbSeq targets are shown.

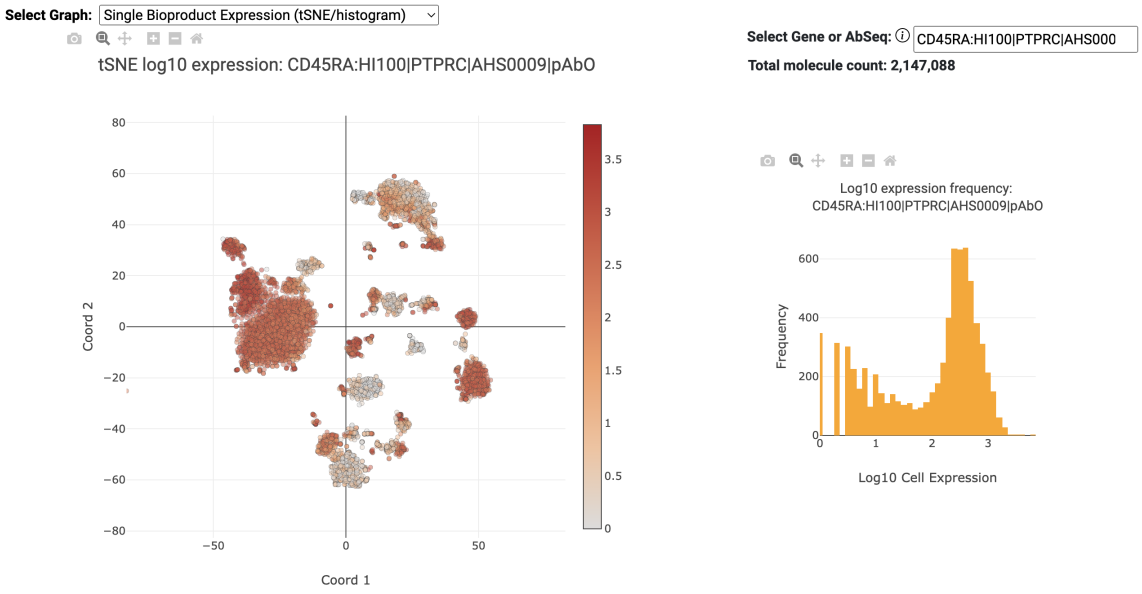


Figure 17 Single Bioproduct Expression tSNE and histogram for the AbSeq target CD45RA

Immune Cell Type Experimental

The Immune Cell Type Experimental graph shows the tSNE plot with each cell labeled according to the results from the cell type prediction algorithm.

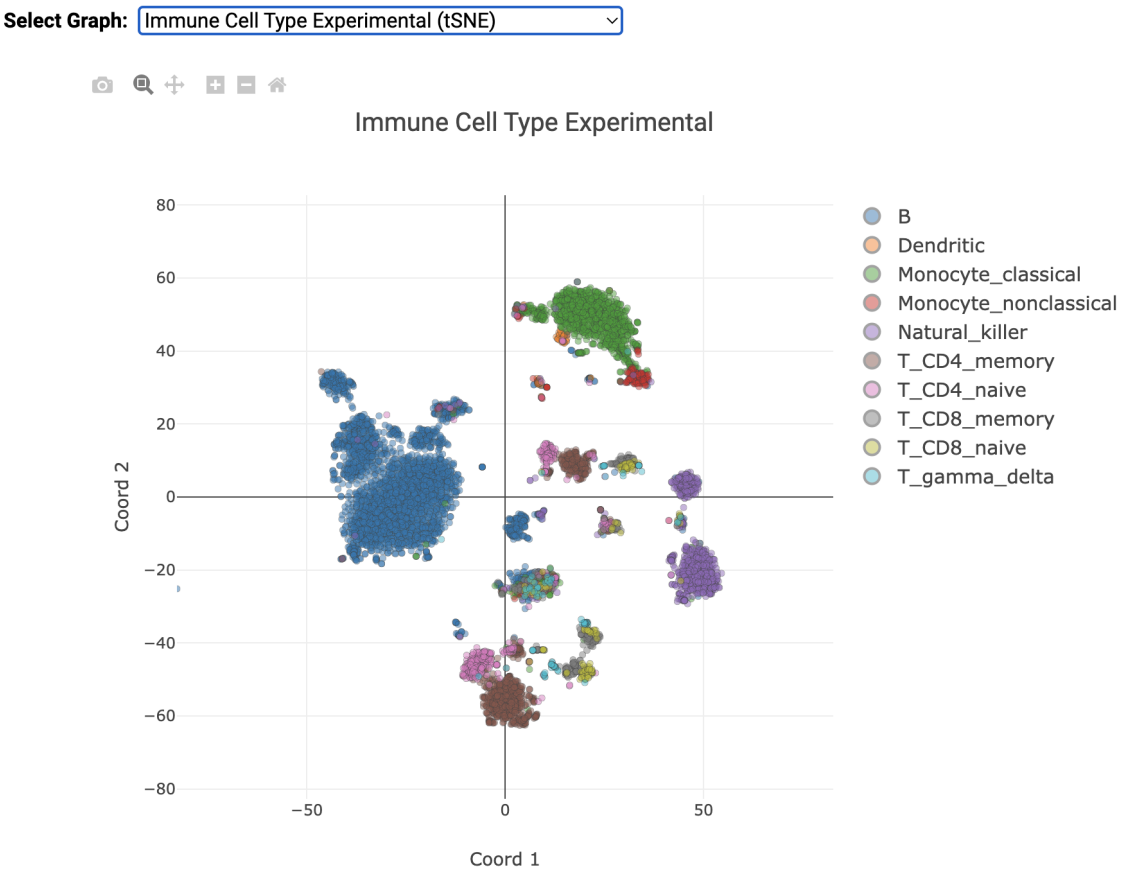


Figure 18 Immune Cell Type Experimental for a VDJ dataset

Total Molecules per cell (mRNA and AbSeq)

The Total Molecules per cell mRNA and AbSeq graphs show the tSNE plot on the left where each cell is colored by the log 10 of total expression for all mRNA genes or AbSeq targets respectively. The histogram on the right shows the distribution of total expression for all cells for all mRNA genes or AbSeq targets respectively.

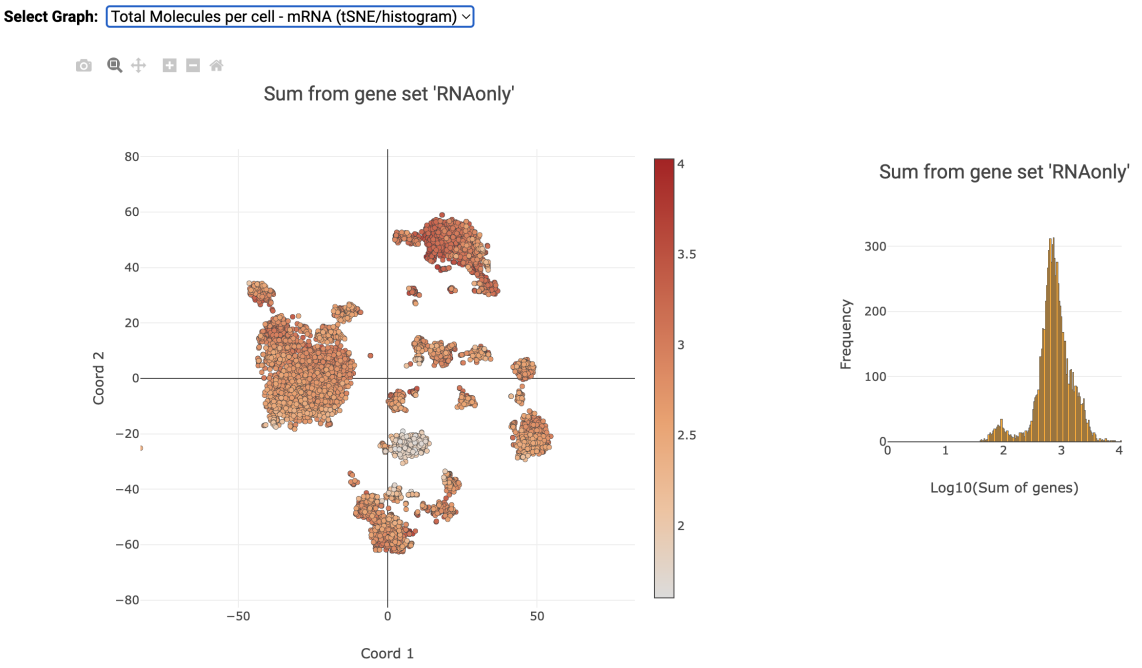


Figure 19 Total Molecules per cell for all mRNA genes

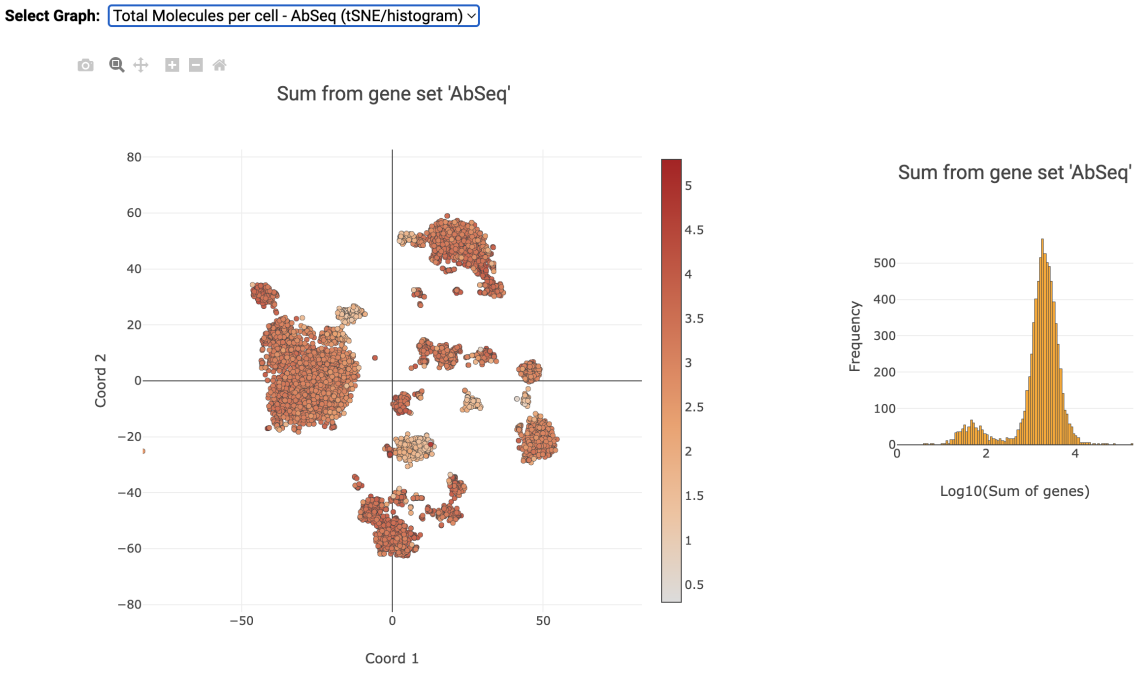


Figure 20 Total Molecules per cell for all AbSeq targets

VDJ BCR/TCR Paired Chains

The VDJ BCR/TCR Paired Chains tSNE plots show the clusters of cells with BCR/TCR paired chains.

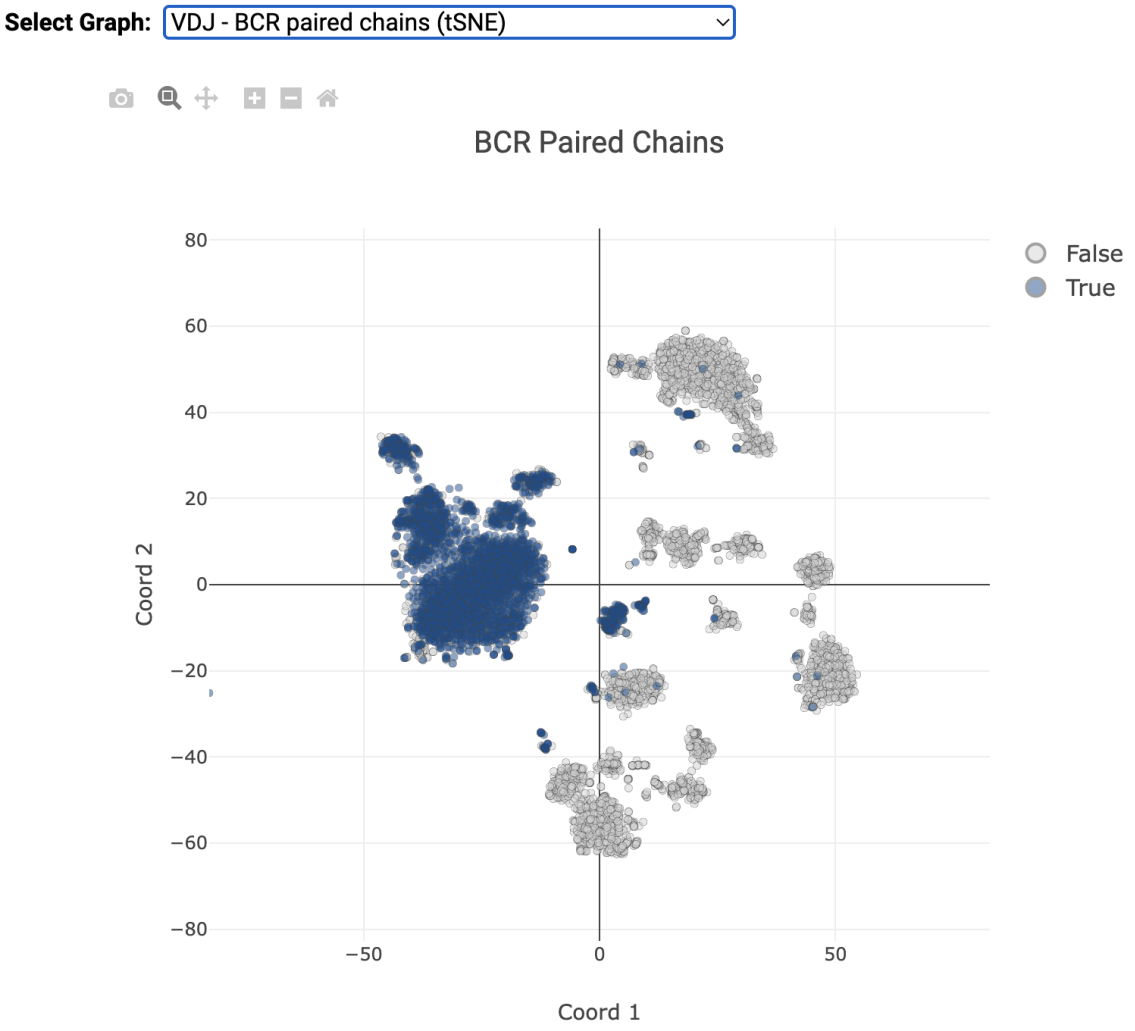


Figure 21 VDJ BCR Paired Chains for a VDJ dataset

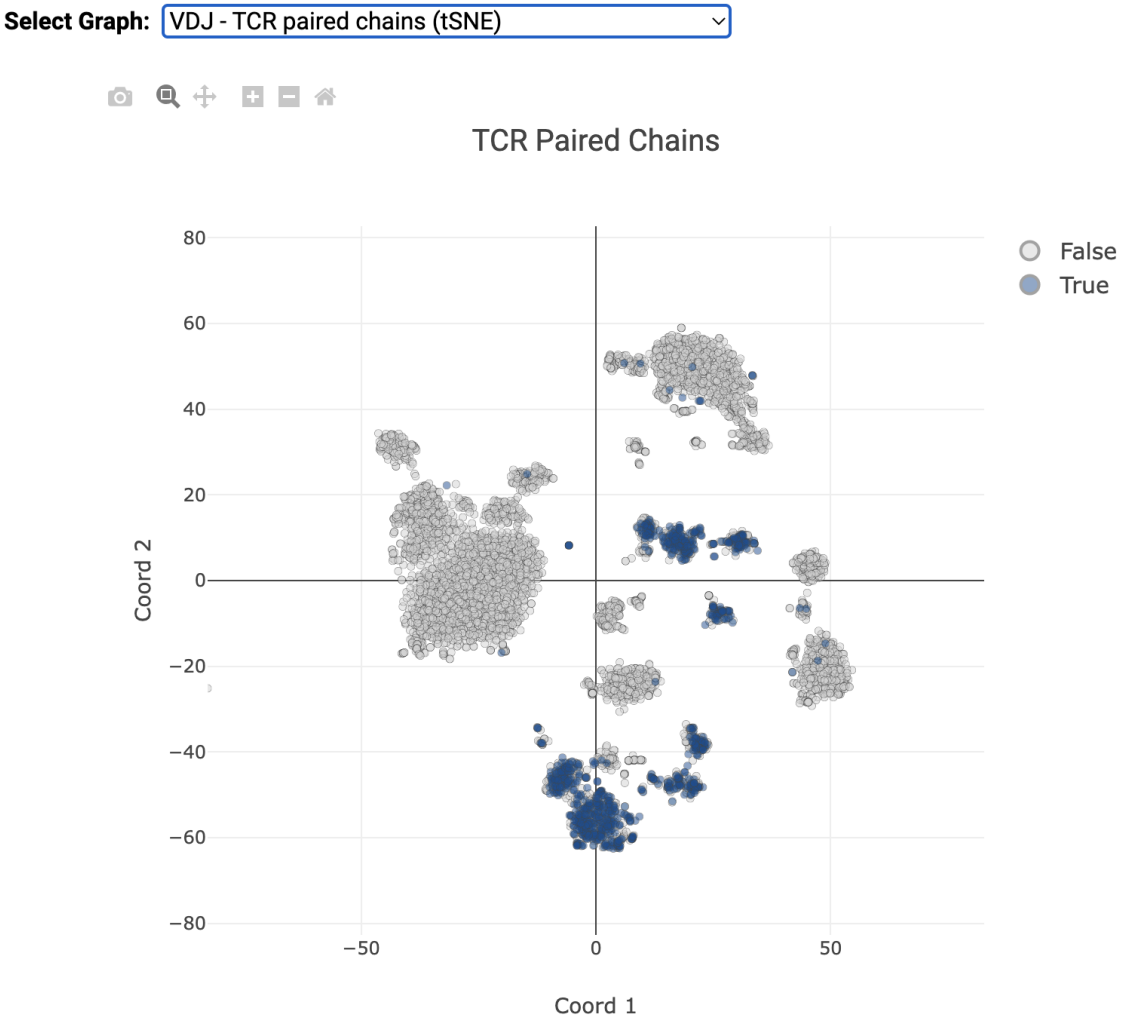


Figure 22 VDJ TCR Paired Chains for a VDJ dataset

Sample Multiplexing

The Sample Multiplexing tSNE plot shows the cells labeled by sample tag and includes the multiplet and undetermined cell labels.

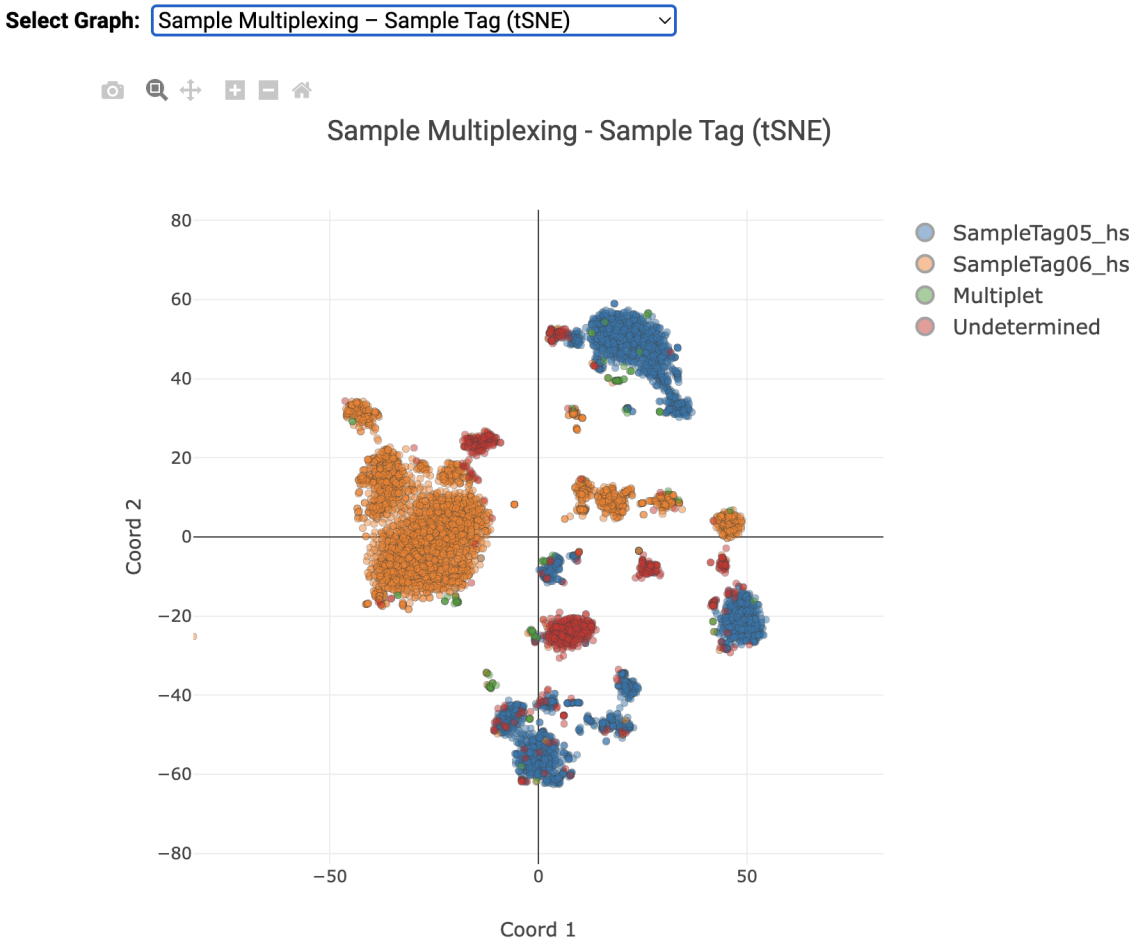


Figure 23 Sample Tag Multiplexing for a VDJ dataset

Metric Sections

There are several sections in the pipeline report providing details about specific metrics. The main sections cover the Sequencing Quality, Library Quality, Alignment Categories, Reads and Molecules, Cell Calling, Error Correction, Sample Multiplexing, and VDJ results. The data in these sections is identical to the Metrics Summary CSV file. More details about some of the sections are provided below.

Cell Calling

The Cell Calling section provides an interactive graph from the basic cell calling algorithm that was described in Step 6. Determine putative cells on page 15. The second derivative plot is shown on top of the cumulative read plot and the basic cell line is shown in red. Hovering over the graph will display a vertical line that corresponds to the number of putative cells on the cumulative read plot. All general graph functionality is available. See General graph functionality on page 34 for details.

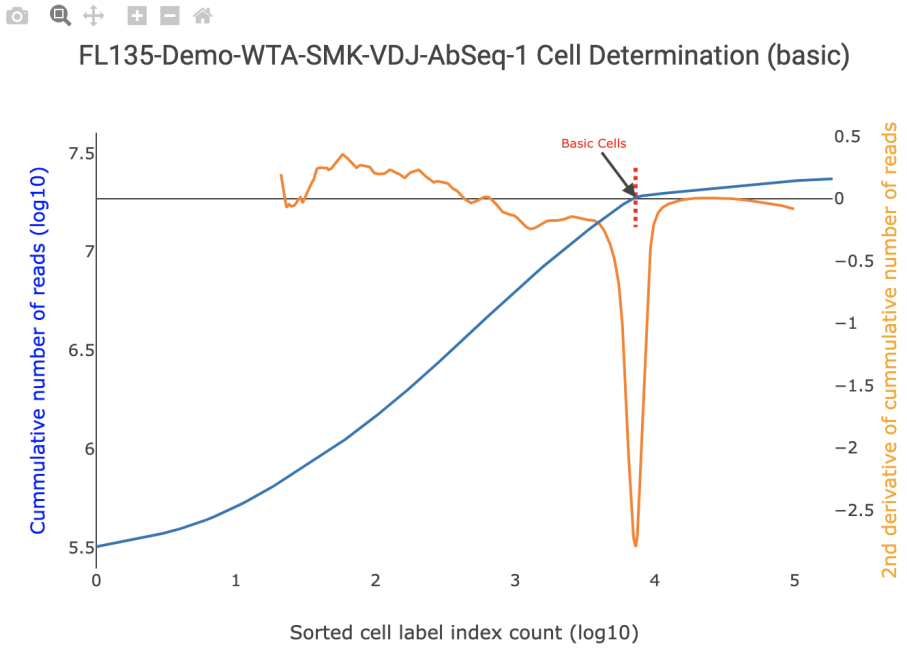


Figure 24 Basic cell calling graph

Sample Multiplexing

In the sample multiplexing section, there is summary information such as the number of filtered reads that aligned to the sample tags and the percentage of sample tag reads that are assigned to putative cells. There is also a detailed section showing the number of reads and percentage of reads assigned to each sample tag, along with the number of cells, percentage of cells, number of reads per cell, and mean reads per cell for each sample tag. The detailed section also shows the number of multiplets and undetermined cells.

VDJ

In the VDJ section, the first table for the “Reads” and the second table for the “Molecules and Dominant Contigs” show the collapsed summary information for the Chain Category (BCR/TCR). By pressing the down arrow, the table expands to show more details about the specific chains. There is also a section for Cell Type

specific metrics. There are four tables that can be selected from the dropdown menu: Paired Chains Pct, Pct Cells Positive, Pct Cells Full Length, and Mean Molecules per Cell.

Metric Alerts

The Metric Alerts section provides information about metrics from the experiment that are above or below certain thresholds that are typical for most experiments. The alert will specify the library or bioproduct, metric, metric value, threshold, and some possible causes and suggestions.

General graph functionality

There are several ways to interact with the graphs. The toolbar provides the following functionality (from left to right):

Graph function	Description
Download Plot	Allows you to download the plot in SVG format. Once downloaded, the SVG is a static image.
Zoom	Allows you to create a box which will zoom in to show the selected region in the graph area.
Pan	Allows you to move the graph to center on a different part of the graph to observe it clearer.
+	Zooms in 1 level around the center of the graph.
–	Zooms out 1 level around the center of the graph.
Home	Resets graph to original zoom and axes.
Additional features	
Color bar	The color bar on the right side of the Single Bioproduct Expression and Total Molecules per cell (mRNA and AbSeq) tSNE plots shows the intensity of log 10 based expression.
Hover	Hovering over the points on the graphs will give extra information (for example, cell index or expression level).

Reviewing sequencing analysis output files

Before you begin

Obtain the output files after running the appropriate pipeline on the Seven Bridges Genomics platform or on a local installation. See the *BD® Single-Cell Multiomics Analysis Setup User Guide* (23-21333).

Sequencing analysis outputs

Most outputs contain a header summarizing the pipeline run. Headers contain all of the information needed to re-run the pipeline with the same settings.

Output	File	Content
Metrics summary on page 37	<sample_name>_Metrics_Summary.csv	Report containing sequencing, molecules, and cell metrics

Output	File	Content
Pipeline report HTML on page 26 (described in Step 11. Generate summary on page 26)	<sample_name>_Pipeline_Report.html	Summary report containing the results from the sequencing analysis pipeline run
BAM and BAM Index on page 45	<sample_name>.BAM <sample_name>.BAM.bai	Alignment file of R2 and associated R1 annotations
Data tables on page 46	<sample_name>_RSEC_MolsPerCell.csv <sample_name>_RSEC_ReadsPerCell.csv <sample_name>_DBEC_MolsPerCell.csv <sample_name>_DBEC_ReadsPerCell.csv	Reads per bioproduct per cell and molecules per bioproduct per cell, based on RSEC or DBEC
	<sample_name>_RSEC_MolsPerCell_Unfiltered.csv.gz <sample_name>_RSEC_ReadsPerCell_Unfiltered.csv.gz <sample_name>_DBEC_MolsPerCell_Unfiltered.csv.gz <sample_name>_DBEC_ReadsPerCell_Unfiltered.csv.gz	Unfiltered tables containing all cell labels of ≥10 reads
Expression data on page 47	<sample_name>_Expression_Data.st	The expression sparse matrix, a table of counts in sparse format
	<sample_name>_Expression_Data_Unfiltered.st.gz	Compressed file containing all cell labels of ≥10 reads
Putative cells origin on page 48	<sample_name>_Putative_Cells_Origin.csv	Algorithm that found the putative cell: basic or refined
Protein Aggregates Experimental on page 48	<sample_name>_Protein_Aggregates_Experimental.csv	Putative cells identified as protein aggregates

Output	File	Content
Bioproduct Statistics on page 49	<sample_name>_Bioproduct_Stats.csv	Metrics from RSEC and DBEC unique molecular identifier adjustment algorithms on a per-bioproduct basis
Sample Tag metrics (sample multiplexing option selected) on page 50	<sample_name>_Sample_Tag_Metrics.csv	Metrics from the sample determination algorithm
Sample Tag calls (sample multiplexing option selected) on page 51	<sample_name>_Sample_Tag_Calls.csv	Assigned Sample Tag for each putative cell
Per sample folder (sample multiplexing option selected) on page 52	<sample_name>_Sample_Tag<number>.zip <sample_name>_Multiplet_and_Undetermined.zip	Data tables metric summary, and expression matrix for a particular sample. **Note**: For putative cells that could not be assigned a specific Sample Tag, a Multiplet_ and _Undetermined.zip file is also output.
VDJ metrics (VDJ option selected) on page 53	<sample_name>_VDJ_Metrics.csv	Overall metrics from the VDJ analysis
VDJ per Cell metrics (VDJ option selected) on page 54	<sample_name>_VDJ_perCell.csv <sample_name>_VDJ_perCell_uncorrected.csv	Cell specific read and molecule counts, VDJ gene segments, CDR3 sequences, paired chains, and cell type
VDJ Dominant Contigs (VDJ option selected) on page 56	<sample_name>_VDJ_Dominant_Contigs_AIRR.csv	Dominant contig for each cell label - chain type combination (putative cells only)

Output	File	Content
VDJ Unfiltered Contigs (VDJ option selected) on page 59	<sample_name>_VDJ_Unfiltered_Contigs_AIRR.csv	All contigs that were assembled and annotated successfully (all cells)

Metrics summary

File: <sample_name>_Metrics_Summary.csv

The Metrics summary provides statistics on sequencing, molecules, cells, and bioproducts.

Sample Tag and AbSeq metrics display only when they are used in an experiment.

Example of a portion of the output for targeted assays:

#Sequencing Quality#									
Total_Reads_in_FASTQ	Pct_Reads_Too_Short	Pct_Reads_Low_Base_Quality	Pct_Reads_High_SNF	Pct_Reads_Filtered_Out	Total_Reads_After_Quality_Filtering	Library			
51392677	0	1.23	0.6	1.74	50498043	BD-Targeted			
#Library Quality#									
Total_Filtered_Reads	Pct_Contaminating_PhiX_Reads_in_Filtered_R2	Pct_Q30_Bases_in_Filtered_R2	Pct_Assigned_to_Cell_Labels	Pct_Cellular_Reads_Aligned_Uniquely_to_Amplicons	Library				
50498043	0	80.75	94.07	89.08	J034FC2ABC				
#Reads and Molecules#									
Aligned_Reads_By_Type	Total_Raw_Molecules	Total_RSEC_Molecules	Total_DBEC_Molecules	Mean_Raw_Sequencing_Depth	Mean_RSEC_Sequencing_Depth	Mean_DBEC_Sequencing_Depth	Sequencing_Saturation	Pct_Cellular_Reads_with_Amplicons_Retained_by_DBEC	Target_Type
4129881	520034	399451	265471	7.94	10.34	14.65	98.06	94.18	mRNA
40856055	13557243	12346297	10585984	3.01	3.31	3.69	92.94	95.69	AbSeq
44985936	14077277	12745748	10851455	3.2	3.53	3.96	93.41	95.55	mRNA + AbSeq
#Cells RSEC#									
Putative_Cell_Count	Pct_Reads_from_Putative_Cells	Mean_Reads_per_Cell	Mean_Molecules_per_Cell	Median_Molecules_per_Cell	Mean_Targets_per_Cell	Median_Targets_per_Cell	Total_Targets_Detected	Target_Type	
762	87.06	4718.68	405.01	262.5	68.29	64	346	mRNA	
762	33.16	17781.23	5344.74	4460.5	19.97	20	20	AbSeq	
762	38.11	22499.92	5749.75	4950	88.27	83.5	366	mRNA + AbSeq	
#Cells DBEC#									
Putative_Cell_Count	Pct_Reads_from_Putative_Cells	Mean_Reads_per_Cell	Mean_Molecules_per_Cell	Median_Molecules_per_Cell	Mean_Targets_per_Cell	Median_Targets_per_Cell	Total_Targets_Detected	Target_Type	
762	88.68	4526.72	304.3	199	58.55	54	346	mRNA	
762	33.28	17074.17	4637.68	3838.5	19.96	20	20	AbSeq	
762	38.29	21600.9	5749.75	4950	88.27	83.5	366	mRNA + AbSeq	
#Targets#									
Number_of_DBEC_and_RSEC_Corrected_Targets	Number_of_RSEC_Corrected_Targets	Number_of_Targets_in_Panel	Target_Type						
325	21	399	mRNA						
12	8	20	AbSeq						

Example of the output for WTA assays:

#Sequencing Quality#								
Total_Reads_in_FASTQ	Pct_Reads_Too_Short	Pct_Reads_Low_Base_Quality	Pct_Reads_High_SNF	Pct_Reads_Filtered_Out	Total_Reads_After_Quality_Filtering	Library		
36508493	0.19	1.78	3.63	5.11	34644306	BD-Demo-WTA-SMK		
#Library Quality#								
Total_Filtered_Reads	Pct_Contaminating_PhiX_Reads_in_Filtered_R2	Pct_Q30_Bases_in_Filtered_R2	Pct_Assigned_to_Cell_Labels	Pct_Cellular_Reads_Aligned_Uniquely_to_Annotated_Transcriptome	Pct_Cellular_Reads_Aligned_Uniquely_to_Other_Genomic_Regions	Pct_Cellular_Reads_Aligned_Not_Unique	Pct_Cellular_Reads_Unaligned	
34644306	0	88.22	82.61	67.31	12.83	2.38	0.09	BD-Demo-WTA-SMK
#Reads and Molecules#								
Aligned_Reads_By_Type	Total_Raw_Molecules	Total_RSEC_Molecules	Mean_Raw_Sequencing_Depth	Mean_RSEC_Sequencing_Depth	Sequencing_Saturation	Target_Type		
19275337	15310719	15229529	1.26	1.27	37.91	mRNA		
#Cells RSEC#								
Putative_Cell_Count	Pct_Reads_from_Putative_Cells	Mean_Reads_per_Cell	Mean_Molecules_per_Cell	Median_Molecules_per_Cell	Mean_Targets_per_Cell	Median_Targets_per_Cell	Total_Targets_Detected	Target_Type
3324	90.69	5258.75	4151.69	4016.5	1746.67	1794.5	19325	mRNA
#Sample_Tags#								
Sample_Tag_Filtered_Reads	ST_Pct_Reads_from_Putative_Cells							
3103340	74.71							

Metrics summary output

Section/metric	Definition	Major contributing factors
Sequencing Quality		
Total_Reads_in_FASTQ	Number of read pairs in the input FASTQ files	• Sequencing amount
Pct_Read_Pair_Overlap	Percentage of read pairs overlapped with each other	• Sequencing quality
Pct_Reads_Too_Short	Percentage of read pairs filtered out due to length of R2 <40 bp	• Sequencing quality
Pct_Reads_Low_Base_Quality	Percentage of reads filtered out due to average base quality score of R1 reads <20 or R2 reads <20	• Sequencing quality
Pct_Reads_High_SNF	Percentage of read pairs filtered out due to single nucleotide frequency ≥55% for R1 or ≥80% for R2	• Sequencing quality
Pct_Reads_Filtered_Out	Percentage of reads removed by the combination of length, quality, and SNF filters	• Sequencing quality
Total_Reads_After_Quality_Filtering	Number of read pairs after length, quality, and SNF filtering	• Sequencing amount • Sequencing run quality • Library quality
Library	Name of library	• Name of library
Library Quality		
Total_Filtered_Reads	Number of read pairs after length, quality, and SNF filtering	• Sequencing amount • Sequencing run quality • Library quality

Metrics summary output (continued)

Section/metric	Definition	Major contributing factors
Pct_Contaminating_PhiX_Reads_in_Filtered_R2	Percentage of read pairs after quality filtering that are aligned to the PhiX control	• Sequencing run quality • Amount of PhiX spiked in
Pct_Q30_Bases_in_Filtered_R2	Percentage of R2 bases with quality score >30, averaged across all read pairs retained after quality filtering	• Sequencing quality
Pct_Assigned_to_Cell_Labels	Percentage of read pairs containing a valid cell label	• Sequencing quality • Library quality
Pct_Cellular_Reads_Aligned_Uniquely	Percentage of read pairs containing a valid cell label and UMI that aligned uniquely	• Sequencing quality • Library quality
Library	Name of library	• Name of library
Alignment Categories		
Cellular_Reads	Number of read pairs containing a valid cell label and UMI	• Sequencing Quality • Library Quality
mRNA_Targeted_Pct (Targeted Only)	Percentage of cellular reads with read 2 aligned to mRNA panel reference	• Sequencing Quality • Library Quality
AbSeq_Pct (Targeted and WTA)	Percentage of cellular reads with read 2 aligned to AbSeq reference	• Sequencing Quality • Library Quality
Sample_Tag_Pct (Targeted and WTA)	Percentage of cellular reads with read 2 aligned to Sample Tag	• Sequencing Quality • Library Quality
Unaligned_Pct (Targeted and WTA)	Percentage of cellular reads with read 2 unaligned to reference	• Sequencing Quality • Library Quality
Annotated_Transcriptome_Pct (WTA Only)	Percentage of cellular reads with read 2 aligned uniquely to gene present in the transcriptome	• Sequencing Quality • Library Quality • Cell Type
Introns_Pct (WTA Only)	Percentage of cellular reads with read 2 aligned uniquely to an intronic region of a gene	• Sequencing Quality • Library Quality • Cell Type
Intergenic_Regions_Pct (WTA Only)	Percentage of cellular reads with read 2 aligned uniquely to an intergenic region	• Sequencing Quality • Library Quality • Cell Type
Antisense_Pct (WTA Only)	Percentage of cellular reads with read 2 aligned uniquely to an antisense strand	• Sequencing Quality • Library Quality • Cell Type
Not_Unique_Pct (WTA Only)	Percentage of cellular reads with read 2 not uniquely aligned	• Sequencing Quality • Library Quality • Cell Type

Metrics summary output (continued)

Section/metric	Definition	Major contributing factors
Ambiguous_Pct (WTA Only)	Percentage of cellular reads with read 2 aligned to region with ambiguous annotation	• Sequencing Quality • Library Quality • Cell Type
No_Feature_Pct (WTA Only)	Percentage of cellular reads with read 2 aligned to non-annotated region	• Sequencing Quality • Library Quality • Cell Type
Library	Name of library	• Name of library
Reads and Molecules		
Aligned_Reads_By_Type	Number of filtered read pairs aligned to bioproduct type	• Sequencing quality • Library quality • Panel compatibility with sample composition
Total_Raw_Molecules	Total number of molecules as defined by the unique combination of cell label, bioproduct identity, and UMI	• Sequencing depth • Panel compatibility with sample composition
Total_RSEC_Molecules[a]	Total number of molecules detected after the RSEC molecular identifier adjustment algorithm	• Sequencing depth • Panel compatibility with sample composition
Total_DBEC_Molecules[a] (Targeted and AbSeq libraries in WTA)	Total number of molecules detected after RSEC and DBEC molecular identifier adjustment algorithms	• Sequencing depth • Panel compatibility with sample composition
Mean_Raw_Sequencing_Depth	Average number of read pairs per molecule before molecular identifier adjustment algorithms	• Sequencing depth
Mean_RSEC_Sequencing_Depth	Average number of read pairs per molecule after the RSEC molecular identifier adjustment algorithm	• Sequencing depth
Mean_DBEC_Sequencing_Depth	Average number of read pairs per molecule after RSEC and DBEC molecular identifier adjustment algorithms	• Sequencing depth
Sequencing_Saturation	Percentage of read pairs representing RSEC-adjusted molecules that are sequenced more than once	• Sequencing depth
Pct_Cellular_Reads_with_Amplicons_Retained_by_DBEC (Targeted only)	Percentage of read pairs with valid cell labels and bioproduct alignment retained after the DBEC molecular adjustment algorithm	• Sequencing depth
Bioproduct_Type	Type of bioproduct in library (mRNA, AbSeq, or mRNA + AbSeq)	• Library composition

Metrics summary output (continued)

Section/metric	Definition	Major contributing factors
Cells RSEC **Note:** Cells RSEC contains the metrics from cell label filtering based on molecule data generated from the RSEC molecular index adjustment algorithm.		
Putative_Cell_Count[b]	Number of cell labels detected by the cell label filtering algorithm	• Number of cells input and captured by cartridge workflow • Bead handling • Panel compatibility with sample composition
Pct_Reads_from_Putative_Cells	Percentage of reads that are assigned to putative cells	• Cell viability • Cartridge workflow performance • Sequencing depth (for DBEC-derived metric only) • Panel compatibility with sample composition
Mean_Reads_per_Cell	Average number of reads representing the molecules detected in each cell	• Sequencing depth • Panel compatibility with sample composition
Median_Reads_per_Cell	Median number of reads representing the molecules detected in each cell.	• Sequencing depth • Panel compatibility with sample composition
Mean_Molecules_per_Cell	Average number of molecules detected per cell label	• Sequencing depth • Panel compatibility with sample composition
Median_Molecules_per_Cell	Median number of molecules detected per cell label	• Sequencing depth • Panel compatibility with sample composition
Mean_Bioproducts_per_Cell	Average number of bioproducts detected per cell label	• Sequencing depth • Panel compatibility with sample composition

Metrics summary output (continued)

Section/metric	Definition	Major contributing factors
Median_Bioproducts_per_Cell	Median number of bioproducts detected per cell label	• Sequencing depth • Panel compatibility with sample composition
Total_Bioproducts_Detected	Number of bioproducts detected from all cells	• Sequencing depth • Panel compatibility with sample composition
Bioproduct_Type	Type of bioproduct in library (mRNA, AbSeq, or mRNA + AbSeq)	• Panel composition
Cells DBEC **Note:** Cells contains the metrics from cell label filtering based on molecule data generated from the RSEC and DBEC molecular index adjustment algorithm.		
Putative_Cell_Count[b]	Number of cell labels detected by the cell label filtering algorithm	• Number of cells input and captured by cartridge workflow • Bead handling • Panel compatibility with sample composition
Pct_Reads_from_Putative_Cells	Percentage of reads that are assigned to putative cells	• Cell viability • Cartridge workflow performance • Sequencing depth (for DBEC-derived metric only) • Panel compatibility with sample composition
Mean_Reads_per_Cell	Average number of reads representing the molecules detected in each cell	• Sequencing depth • Panel compatibility with sample composition
Median_Reads_per_Cell	Median number of reads representing the molecules detected in each cell	• Sequencing depth • Panel compatibility with sample composition
Mean_Molecules_per_Cell	Average number of molecules detected per cell label	• Sequencing depth • Panel compatibility with sample composition

Metrics summary output (continued)

Section/metric	Definition	Major contributing factors
Median_Molecules_per_Cell	Median number of molecules detected per cell label	• Sequencing depth • Panel compatibility with sample composition
Mean_Bioproducts_per_Cell	Average number of bioproducts detected per cell label	• Sequencing depth • Panel compatibility with sample composition
Median_Bioproducts_per_Cell	Median number of bioproducts detected per cell label	• Sequencing depth • Panel compatibility with sample composition
Total_Bioproducts_Detected	Number of bioproducts detected from all cells	• Sequencing depth • Panel compatibility with sample composition
Bioproduct_Type	Type of bioproduct in library (mRNA, AbSeq, or mRNA + AbSeq)	• Panel composition
Error Correction Level (Targeted and WTA with AbSeq Libraries)		
Number_of_DBEC_and_RSEC_ Corrected	Number of bioproducts with pass status: the bioproducts have sufficient sequencing depth to be considered for adjustment by the DBEC molecular identifier algorithm	• Sequencing depth • Panel compatibility with sample composition
Number_of_RSEC_Corrected	Number of bioproducts not having sufficient sequencing depth to be considered for adjustment by the DBEC molecular identifier algorithm	• Sequencing depth • Panel compatibility with sample composition
Number_in_Panel	The number of bioproducts featured in the panel	• Panel choice
Bioproduct_Type	Type of bioproduct in library (mRNA, AbSeq, or mRNA + AbSeq)	• Library composition
Sample Tags (If used in the experiment)		
Sample_Tag_Filtered_Reads	Number of filtered read pairs aligned to Sample Tags	• Sequencing depth • Panel compatibility with sample composition

Metrics summary output (continued)

Section/metric	Definition	Major contributing factors
ST_Pct_Reads_from_Putative_Cells	Percentage of Sample Tag reads that are assigned to putative cells	• Cell viability • Sample Tag labelling and wash protocols • Cartridge workflow performance • Sequencing depth (for DBEC-derived metric only) • Panel compatibility with sample composition

a. For more information on RSEC and DBEC molecular identifier adjustment algorithms, see Step 5. Annotate molecules on page 11.

b. For further information on how putative cells are defined in terms of the number of reads associated with true and noise cell labels, see Reviewing sequencing analysis output files on page 34.

BAM and BAM Index

BAM File: <sample_name>.BAM
BAM Index: <sample_name>.BAM.bai

BAM is an alignment file in binary format that is generated by the aligner. The aligner aligns R2 reads to the reference file and outputs tags related to alignment quality. This BAM file is sorted according to the alignment coordinates of R2 reads on each chromosome.

The BAM Index is the index file associated with the coordinate-sorted BAM file.

The BD Rhapsody™ Analysis pipeline adds the following tags:

Tag	Definition
CB	A number between 1 and 96^3 (884,736) representing a unique cell label sequence (CB = 0 when no cell label sequence is detected).
MR	Raw molecular identifier sequence.
MA	RSEC-adjusted molecular identifier sequence. If not a true cell, the raw UMI is repeated in this tag.
PT	T if a poly(T) tail was found in the expected position on R1, or F if poly(T) was not found.
CN	Indicates if a sequence is derived from a putative cell, as determined by the cell label filtering algorithm (*T*: putative cell; *x*: invalid cell label or noise cell). **Note:** You can distinguish between an invalid cell label and a noise cell with the CB tag (invalid cell labels are 0).
ST	The value is 1–12, indicating the Sample Tag of the called putative cell, or *M* for multiplet, or *x* for undetermined.
TR (WTA only)	Transcripts associated with the unique alignment. Transcripts are separated by "|".
TF (WTA only)	Mean fragment length based on associated transcripts in TR tag. For transcripts with fragment lengths less than 1000 bp, only values less than 1000 bp are used in calculation of mean.

Note: A BAM file can be converted to a tab-delimited text file (SAM format) by using SAMtools (see samtools.sourceforge.net).

Data tables

Files containing filtered data:

<sample_name>_RSEC_MolsPerCell.csv

<sample_name>_RSEC_ReadsPerCell.csv

<sample_name>_DBEC_MolsPerCell.csv

<sample_name>_DBEC_ReadsPerCell.csv

Compressed files containing unfiltered data:

<sample_name>_RSEC_MolsPerCell_Unfiltered.csv.gz

<sample_name>_RSEC_ReadsPerCell_Unfiltered.csv.gz

<sample_name>_DBEC_MolsPerCell_Unfiltered.csv.gz

<sample_name>_DBEC_ReadsPerCell_Unfiltered.csv.gz

Eight Data Table .csv files, four filtered and four unfiltered, are output. They contain reads per bioproduct per cell and molecules per bioproduct per cell.

For example:

Cell_Index	ADA	ADGRE1	ADGRG3	ADM	AIM2	ALAS2	ANXA5	AOC3
525435	5	0	0	0	0	0	0	0
268870	3	0	0	0	0	0	0	0
38817	22	0	0	0	0	0	0	0
24642	19	0	0	0	0	0	1	0
444017	5	0	0	0	0	0	0	0
771197	2	0	0	0	0	0	0	0
480465	8	0	0	0	0	0	1	0
161815	0	0	0	0	0	0	0	0
379509	2	0	0	0	0	0	0	0
757154	3	0	0	0	0	0	0	0
25539	4	0	0	0	0	0	0	0
548867	2	0	0	0	0	0	0	0
297014	0	0	0	0	0	0	0	0
714491	1	0	0	0	0	0	0	0
604203	0	0	0	0	0	0	0	0

- Each row represents the number of reads or molecules in a cell for each bioproduct in the panel (targeted) or bioproduct detected (WTA). A cell is identified with a unique cell index number under Cell_Index.
- The cell index is sorted in descending order based on the total number of reads. The cell order in the four files is the same.
- Bioproduct names are sorted by bioproduct type (AbSeq followed by mRNA). Within each bioproduct type, bioproduct names are sorted alphabetically.
- For PerCell.csv files: Reads and molecules are counted only if they have passed all pipeline filters and have been determined to be from putative cells.
- For PerCell_Unfiltered.csv.gz: The files contain unfiltered tables with cell labels of ≥10 reads.

Note: It is generally recommended to use <sample_name>_DBEC_MolsPerCell.csv for clustering analysis. Read counts for DBEC, read counts for RSEC, and molecule counts for RSEC are provided for reference. The RSEC files can be used when sequencing depth is so low that most bioproducts do not pass the threshold for the DBEC molecular identifier adjustment algorithm to be applied—that is, low_depth in <sample_name>_Bioproduct_Stats.csv.

Expression data

File: <sample_name>_Expression_Data.st

Unfiltered file: <sample_name>_Expression_Data_Unfiltered.st.gz

Information is presented in sparse notation.

- Data.st: Reads and molecules are counted only if they have passed all pipeline filters and have been determined to be from putative cells.
- Unfiltered.st.gz: Compressed file containing all cell labels of ≥10 reads.

Open the .st file in a text editor.

Each row records counts for cell-bioproduct combinations that have non-zero RSEC molecule counts.

For example:

Cell_Index	Bioproduct	RSEC_Reads	Raw_Molecules	RSEC_Adjusted_Molecules	DBEC_Reads	DBEC_Adjusted_Molecules
109559	IGHA1_secreted	1	1	1	1	
109559	IGHG2_secreted	5	5	5	5	
109559	IGKC	28	28	28	28	
109559	IL1B	1	1	1	1	
693695	ADA	3	3	3	3	
693695	CD3D	1	1	1	1	
693695	CHI3L2	1	1	1	1	
693695	ITGB2	1	1	1	1	
693695	PCNA	1	1	1	1	
134770	ADA	1	1	1	1	
134770	CD3D	1	1	1	1	
134770	FYB	1	1	1	1	
134770	GAPDH	1	1	1	1	
134770	LCK	1	1	1	1	
134770	PTTG2	1	1	1	1	
134770	SELL	1	1	1	1	
28621	CD3D	2	2	2	2	
28621	CD3|CD3E|AHS0033|pAbO	1	1	1	1	1
28621	CD45|PTPRC|AHS0040|pAbO	1	1	1	1	1
28621	GAPDH	2	2	2	2	
28621	IKZF2	1	1	1	1	
28621	ITGAE	1	1	1	1	
28621	RUNX3	1	1	1	1	
635149	CD74	1	1	1	1	
635149	FCN1	1	1	1	1	
635149	GAPDH	1	1	1	1	
635149	IL15RA	1	1	1	1	
635149	S100A12	1	1	1	1	
563518	AURKB	1	1	1	1	

Metric	Definition
Cell_Index	Unique cell index sorted by total number of reads per cell in descending order
Bioproduct	Bioproduct names listed in alphabetical order per cell index
RSEC_Reads	Number of reads after the RSEC molecular identifier adjustment algorithm
Raw_Molecules	Number of UMIs before molecular identifier adjustment algorithms
RSEC_Adjusted_Molecules	Number of UMIs after RSEC molecular identifier adjustment algorithm
DBEC_Reads	Number of reads remaining after the DBEC molecular identifier adjustment algorithm
DBEC_Adjusted_Molecules	Number of UMIs after RSEC and DBEC molecular identifier adjustment algorithms

Putative cells origin

File: <sample_name>_Putative_Cells_Origin.csv

The output lists the step in the cell label filtering algorithm that determined a particular cell is a putative cell. If the cell label is categorized as putative in the basic implementation of the second derivative analysis, it is labeled *Basic*. If the cell label is a recovered false negative in the refined implementation, it is labeled *Refined*. See Step 6. Determine putative cells on page 15. For example:

##########	
## BD Targeted Rhapsody Analysis Pipeline Version 1.0	
## Analysis Date: 2017-08-03 23:57:20	
## Sample: mySample	
## Reference: Immune_Response_Panel_Hs	
##########	
Cell_Index	Algorithm
525435	Basic
268870	Basic
38817	Basic
24642	Basic
444017	Basic
771197	Basic
480465	Basic
161815	Basic
379509	Basic
757154	Basic
25539	Basic
548867	Basic

Protein Aggregates Experimental

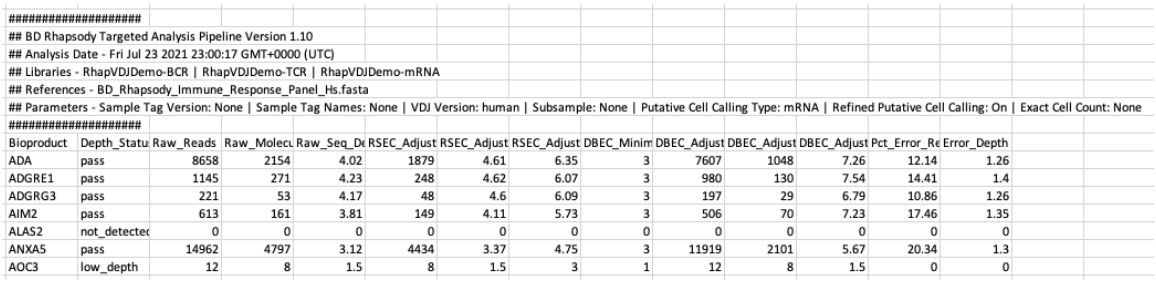
File: <sample_name>_Protein_Aggregates_Experimental.csv

This output only exists when putative cell calling was done using the AbSeq read counts for Putative Cell Identification. Cell labels that are considered to be protein aggregates are set to True.

Bioproduct Statistics

File: <sample_name>_Bioproduct_Stats.csv

The molecular identifier adjustment algorithms RSEC and DBEC are applied to each bioproduct. The molecular identifier metrics file lists the metrics from RSEC and DBEC on a per-bioproduct basis. For more information on RSEC and DBEC molecular identifier adjustment algorithms, see Step 5. Annotate molecules on page 11. For example:



```
####################
## BD Rhapsody Targeted Analysis Pipeline Version 1.10
## Analysis Date - Fri Jul 23 2021 23:00:17 GMT+0000 (UTC)
## Libraries - RhapVDJDemo-BCR | RhapVDJDemo-TCR | RhapVDJDemo-mRNA
## References - BD_Rhapsody_Immune_Response_Panel_Hs.fasta
## Parameters - Sample Tag Version: None | Sample Tag Names: None | VDJ Version: human | Subsample: None | Putative Cell Calling Type: mRNA | Refined Putative Cell Calling: On | Exact Cell Count: None
####################
```

Bioproduct	Depth_Statu	Raw_Reads	Raw_Molecu	Raw_Seq_De	RSEC_Adjust	RSEC_Adjust	RSEC_Adjust	DBEC_Minim	DBEC_Adjust	DBEC_Adjust	DBEC_Adjust	Pct_Error_Re	Error_Depth
ADA	pass	8658	2154	4.02	1879	4.61	6.35	3	7607	1048	7.26	12.14	1.26
ADGRE1	pass	1145	271	4.23	248	4.62	6.07	3	980	130	7.54	14.41	1.4
ADGRG3	pass	221	53	4.17	48	4.6	6.09	3	197	29	6.79	10.86	1.26
AIM2	pass	613	161	3.81	149	4.11	5.73	3	506	70	7.23	17.46	1.35
ALAS2	not_detectec	0	0	0	0	0	0	0	0	0	0	0	0
ANXA5	pass	14962	4797	3.12	4434	3.37	4.75	3	11919	2101	5.67	20.34	1.3
AOC3	low_depth	12	8	1.5	8	1.5	3	1	12	8	1.5	0	0

Metric	Definition
Bioproduct	Bioproduct names listed in alphabetical order
Status	Bioproduct status across all reads and molecules: • Not detected: Bioproduct was not detected, because it has zero reads • Low depth: Minimum sequencing depth not achieved • Pass: Minimum sequencing depth has been achieved
Raw_Reads	Number of reads before molecular identifier adjustment algorithms
Raw_Molecules	Number of UMIs before molecular identifier adjustment algorithms
Raw_Seq_Depth	Number of raw reads ÷ the number of raw molecules
RSEC_Adjusted_Molecules	Number of molecules detected after RSEC molecular identifier adjustment algorithm
RSEC_Adjusted_Seq_Depth	Number of raw reads ÷ the number of RSEC-adjusted molecules
RSEC_Adjusted_Seq_Depth_ without_Singletons	Number of raw reads ÷ the number of RSEC-adjusted molecules without considering molecules represented by only one read
DBEC_Minimum_Depth	Threshold of RSEC depth for a molecule to be considered a putative molecule by DBEC
DBEC_Adjusted_Reads	Number of reads retained after DBEC molecular identifier adjustment algorithm
DBEC_Adjusted_Molecules	Number of molecules retained after RSEC and DBEC
DBEC_Adjusted_Seq_Depth	Number of DBEC-adjusted reads ÷ the number of molecules detected after RSEC and DBEC
Pct_Error_Reads	Percentage of reads removed by DBEC molecular identifier adjustment algorithm
Error_Depth	RSEC depth of molecules that are removed by DBEC correction

Sample Tag metrics (sample multiplexing option selected)

File: <sample_name>_Sample_Tag_Metrics.csv

The Sample Tag metrics file contains statistics on the reads aligned to each Sample Tag and cells called for each sample. For example:

#####################							
## BD Targeted Multiplex Rhapsody Analysis Pipeline Version 1.01							
## Analysis Date: 2017-10-27 08:07:06							
## Sample: T26FC1NB							
## Reference: onco_bc_panel_hs_with_phix							
## Sample Tags Version: Hs							
#####################							
Sample_Tag	Sample_Nam	Raw_Reads	Pct_of_Raw_Reads	Cells_Called	Pct_of_Putative_Ce	Raw_Reads_in_Called	Mean_Reads_per_
All_Tags		16163862	100	1787	100	0	0
SampleTag01_hs	Jurkat_1	2938864	18.18	262	14.66	1616700	6170.61
SampleTag02_hs	Jurkat_2	3928186	24.3	273	15.28	2175688	7969.55
SampleTag03_hs	Ramos_1	4052350	25.07	265	14.83	1997990	7539.58
SampleTag04_hs	Ramos_2	4171232	25.81	278	15.56	2126098	7647.83
SampleTag05_hs	T47D_1	484744	3	356	19.92	315126	885.19
SampleTag06_hs	T47D_2	588480	3.64	291	16.28	377908	1298.65
Multiplet		0	0	59	3.3	0	0
Undetermined		0	0	3	0.17	0	0

File	Description	Major contributing factors
Sample_Tag	List of the Sample Tags in the pipeline run	—
Sample_Name	User-provided sample name	—
Raw_Reads	Number of reads aligned to each Sample Tag	• Sample Tag sequencing amount
Pct_of_Raw_Reads	Percentage of Sample Tag reads aligned to each Sample Tag	• Sample Tag sequencing amount
Cells_Called	Number of putative cells called for each Sample Tag	• Number of cells input and captured by cartridge workflow • Sample Tag sequencing amount
Pct_of_Putative_Cells_Called	Percentage of putative cells called for each Sample Tag	• Number of cells input and captured by cartridge workflow • Sample Tag sequencing amount
Raw_Reads_in_Called_Cells	Number of Sample Tag reads that are assigned to called cells	• Sample Tag sequencing amount
Mean_Reads_per_Called_Cell	Average number of Sample Tag reads representing each called cell	• Sample Tag sequencing amount

Sample Tag calls (sample multiplexing option selected)

File: <sample_name>_Sample_Tag_Calls.csv

The Sample Tag calls file contains the determined sample call for every putative cell. Sample names that you provided are included in a separate column. The Sample Tag calls file can be used to annotate the main data tables, which contain results from all samples. For example:

#####################		
## BD Targeted Multiplex Rhapsody Analysis Pipeline Version 1.01		
## Analysis Date: 2017-10-27 08:07:06		
## Sample: T26FC1NB		
## Reference: onco_bc_panel_hs_with_phix		
## Sample Tags Version: Hs		
#####################		
Cell_Index	Sample_Tag	Sample_Name
205097	SampleTag05_hs	T47D_1
165394	SampleTag05_hs	T47D_1
855569	SampleTag01_hs	Jurkat_1
249537	SampleTag03_hs	Ramos_1
323327	SampleTag04_hs	Ramos_2
696623	Multiplet	Multiplet
635228	SampleTag05_hs	T47D_1
314225	SampleTag02_hs	Jurkat_2
4570	SampleTag01_hs	Jurkat_1
570473	Undetermined	Undetermined
199238	SampleTag02_hs	Jurkat_2
293711	SampleTag03_hs	Ramos_1

File	Description
Cell_Index	Unique cell identifier
Sample_Tag	List of the Sample Tags in the pipeline run
Sample_Name	User-provided sample name

Per sample folder (sample multiplexing option selected)

File: <sample_name> _Sample_Tag<number>.zip

or <sample_name>_Multiplet_and_Undetermined.zip

Either zipped file includes:

<sample_name> _Sample_Tag<number>_DBEC_MolsPerCell.csv

<sample_name> _Sample_Tag<number>_DBEC_ReadsPerCell.csv

<sample_name> _Sample_Tag<number>_RSEC_MolsPerCell.csv

<sample_name> _Sample_Tag<number>_RSEC_ReadsPerCell.csv

<sample_name> _Sample_Tag<number>_Expression_Data.st

<sample_name>_Sample_tag<number>_Metric_Summary.csv

Each sample with at least one called putative cell will generate a sample-specific folder containing data tables. The formats of the files are the same as described in Data tables on page 46.

Data for putative cells that could not be assigned to a specific sample are found in the Multiplet and Undetermined folder.

TCR and BCR output files

Results from TCR and BCR analysis are output in the following files:

VDJ metrics (VDJ option selected)

(*runName*)_VDJ_metrics.csv

Metrics specific to TCR and BCR data, also broken down by chain and by cell type (experimental).

Overall VDJ Metrics

VDJ metrics output

Section/metric	Definition	Major contributing factors
Reads_Cellular_Aligned_to_VDJ	Number of reads with a valid cell label and UMI that aligned to a VDJ gene segment per chain category.	• Sequencing quality • Library quality
Reads_Contig_Assembled	Number of cellular VDJ aligned reads that were assembled into a contig.	• Cell viability • Library quality
Reads_VDJ_Annotated	Number of reads in contigs passing e-value quality filter.	• Cell viability • Library quality
Reads_Putative	Number of Reads_VDJ_Annotated that came from a putative cell.	• Cartridge workflow performance
Reads_Corrected	Number of putative VDJ reads that are from dominant contigs and remain after distribution-based error correction.	• Cell viability • Library quality
Pct_Reads_Corrected	Percent reads of the above metric relative to Reads_ Contig_Assembled.	• Cell viability • Library quality
Mean_Reads_Corrected_per_ Putative_Cell	Average corrected reads per putative cell.	• Cell viability • Library quality
Molecules_VDJ_Annotated	Number of molecules represented by reads in Reads_ VDJ_Annotated metric.	• Cell viability • Library quality
Molecules_Corrected	Number of molecules represented by reads in Reads_ Corrected metric.	• Cell viability • Library quality
Mean_Molecules_Corrected_per_ Putative_Cell	Average number of molecules per putative cell (Molecules_Corrected_Putative / num putative cells).	• Cell viability • Library quality
Dominant_Contigs_Mean_ Nucleotide_Length	Average protein-coding nucleotide length for all dominant contigs.	• Library quality
Dominant_Contigs_Pct_Full_Length	Percent of dominant contigs from putative cells that are VDJ full length contigs.	• Library quality
Dominant_Contigs_Pct_With_CDR3	Percent of dominant contigs from putative cells with CDR3.	• Library quality
Chain_Category	Category for chains such as BCR and TCR.	• VDJ recombination

Chain type metrics

Chain type metrics are identical to overall metrics except that they are split by VDJ chain type, such as TCR Alpha and BCR Kappa.

Cell type metrics

Cell type metrics output

Section/metric	Definition	Major contributing factors
Cell_Type_Experimental	Inferred cell type. Cell type is inferred, either from the mRNA targeted panel expression data or from relative counts of BCR vs TCR.	• Sample type • mRNA panel
Number_cells	Number of cells classified as this cell type.	• Sample type
BCR_Paired_Chains_Pct_Any	Percent of cells of each type that had both a BCR heavy chain and BCR light chain (Kappa or Lambda).	• Cell viability • Library quality
TCR_Paired_Chains_Pct_Any	Percent of cells of each type that had either TCR Alpha and TCR Beta, or TCR Gamma and TCR Delta.	• Cell viability • Library quality
BCR_Paired_Chains_Pct_Full	Percent of cells of each type that had full-length contigs for both BCR heavy chain and BCR light chain (Kappa or Lambda).	• Cell viability • Library quality
TCR_Paired_Chains_Pct_Full	Percent of cells of each type that had full-length contigs for either TCR Alpha and TCR Beta, or TCR Gamma and TCR Delta.	• Cell viability • Library quality
<chain_type>_Pct_Cells_Positive	Percent of cells of each cell type that had at least one valid corrected contig of the listed chain type.	• Cell viability • Library quality
<chain_type>_Pct_Cells_Full_Length	Percentage of cells from each cell type which had a full length contig with the listed chain type.	• Cell viability • Library quality
<chain_type>_Mean_Molecules_per_Cell	Mean number of corrected molecules of the listed chain type in each cell type.	• Cell viability • Library quality

VDJ per Cell metrics (VDJ option selected)

***(runName)*_VDJ_perCell.csv**

Putative cells only, cell order is the same as gene expression ([RSEC/DBEC]_MolsPerCell.csv file).
Only dominant contigs, all error correction applied.
Data columns: Read and molecule counts, VDJ gene segments, CDR3 sequence, pairing, and cell type.

VDJ per cell output

Section/metric	Definition	Major contributing factors
Cell_Index	Unique cell ID for the cell represented by this row. Cell index will match between VDJ data and gene/AbSeq expression data tables.	• Sequencing quality • Library quality

VDJ per cell output (continued)

Section/metric	Definition	Major contributing factors
Total_VDJ_Read_Count	Total number of error-corrected VDJ reads for all chains in the cell.	• Cell viability • Library quality
Total_VDJ_Molecule_Count	Total number of error-corrected VDJ molecules for all chains in the cell.	• Cell viability • Library quality
<chain_type>_V_gene_Dominant	Dominant V gene segment identified for this chain type in the cell.	• VDJ recombination
<chain_type>_D_gene_Dominant	Dominant D gene segment identified for this chain type in the cell.	• VDJ recombination
<chain_type>_J_gene_Dominant	Dominant J gene segment identified for this chain type in the cell.	• VDJ recombination
<chain_type>_C_gene_Dominant	Dominant C gene segment identified for this chain type in the cell.	• VDJ recombination
<chain_type>_CDR3_Nucleotide_Dominant	Nucleotide sequence of the dominant clone for this chain type in the cell.	• VDJ recombination
<chain_type>_CDR3_Translation_Dominant	Amino acid sequence of the dominant clone for this chain type in the cell.	• VDJ recombination
<chain_type>_Read_Count	Number of error-corrected reads for this chain type in the cell.	• Cell viability • Library quality
<chain_type>_Molecule_Count	Number of unique error-corrected molecules (UMI) for this chain type in the cell.	• Cell viability • Library quality
BCR_Paired_Chains	True/False—this cell contains at least one error-corrected molecule of each BCR heavy and light (Kappa or Lambda).	• Cell viability • Library quality
TCR_Paired_Chains	True/False—this cell contains at least one error-corrected molecule of each TCR Alpha and TCR Beta, or TCR Gamma and TCR Delta.	• Cell viability • Library quality
Cell_Type_Experimental	Inferred cell type of this cell index. Cell type is inferred, either from the mRNA targeted panel expression data or from relative counts of BCR vs TCR.	• Sample type • mRNA panel

(*runName*)_VDJ_perCell_uncorrected.csv

All cell IDs – putative and non-putative.
Only dominant contigs, no error correction, no chain family consolidation.
Data columns: Read and molecule counts, VDJ gene segments, CDR3 sequence, pairing, and cell type.

Shared column definitions are identical to the VDJ_perCell.csv file.

VDJ per cell uncorrected output

Section/metric	Definition	Major contributing factors
Putative_Cell	True/False—this cell index was selected as a putative cell based on the mRNA Panel.	• Cell viability • mRNA panel

VDJ Dominant Contigs (VDJ option selected)

***(runName)*_VDJ_Dominant_Contigs_AIRR.tsv**

Putative cells only, dominant contig for each cell ID – chain combination. DBEC adjustment is applied. The file is compliant with the AIRR rearrangement schema and contains additional informational columns in addition to all the mandatory ones.

Data columns: Cell Identifiers, Read and Molecule counts, Full trimmed contig nucleotide and amino acid sequence, Framework and CDR region nucleotide and amino acid sequence, V, D, J, and C gene segments, full length and productive status.

Refer to docs.airr-community.org/en/stable/datarep/rearrangements.html#.

VDJ dominant contigs AIRR output

Section/metric	Definition	Major contributing factors
cell_id	Unique cell ID for the cell represented by this row. Cell index will match between VDJ data and gene/AbSeq expression data tables.	• Sequencing quality • Library quality
cell_type_experimental	Inferred cell type. Cell type is inferred, either from mRNA targeted panel expression data or from relative counts of BCR vs TCR.	• Sample type • mRNA panel
locus	Type of VDJ sequence: one of TRA, TRB, TRG, TRD, IGH, IGK, and IGL.	• Cell viability • Library quality
sequence_id	Unique ID for contig formatted as <cell_id_<locus>_Number.	• Sequencing quality • Library quality
consensus_count	Number of reads for this contig.	• Cell viability • Library quality
duplicate_count	Number of unique molecules (UMI) for this contig.	• Cell viability • Library quality
sequence	Assembled nucleotide sequence of contig after trimming.	• Library quality • VDJ recombination
sequence_length	Length of full contig nucleotide sequence after trimming.	• Library quality • VDJ recombination
sequence_aa	Amino acid sequence of contig after trimming.	• Library quality • VDJ recombination
sequence_aa_length	Length of full contig amino acid sequence after trimming.	• Library quality • VDJ recombination

VDJ dominant contigs AIRR output (continued)

Section/metric	Definition	Major contributing factors
sequence_alignment	Nucleotide sequence corresponding to VDJ coding region after trimming.	• VDJ recombination
sequence_alignment_length	Length of nucleotide sequence corresponding to VDJ coding region after trimming.	• VDJ recombination
sequence_alignment_aa	Amino acid sequence corresponding to VDJ coding region after trimming.	• VDJ recombination
sequence_alignment_aa_length	Length of amino acid sequence corresponding to VDJ coding region after trimming.	• VDJ recombination
germline_alignment	Assembled, aligned, full-length inferred germline sequence spanning the same region as the sequence_ alignment field.	• VDJ recombination
junction	Junction region nucleotide sequence, where the junction is defined as the CDR3 plus the two flanking conserved codons.	• VDJ recombination
junction_aa	Amino acid translation of the junction.	• VDJ recombination
productive	True/False—this cell chain combination contains some amino acid sequence for each framework (FR1-FR4) region and each CDR (1-3) region.	• VDJ recombination
rev_comp	True/False—the alignment is on the opposite strand (reverse complemented) with respect to the contig sequence.This field is always False for contig sequences from the BD Rhapsody VDJ library.	• Sequencing quality • Library quality
complete_vdj	True/False—this cell chain combination contains some amino acid sequence for each framework (FR1-FR4) region and each CDR (1-3) region.	• Library quality • VDJ recombination
v_call	V gene segment identified for this contig.	• VDJ recombination
v_support	Quality of V gene alignment - lower is better.	• Sequencing quality • Library quality
v_cigar	CIGAR string for the V gene alignment.	• VDJ recombination
v_sequence_start	Start position of the V gene in the contig sequence (1-based closed interval).	• VDJ recombination
v_sequence_end	End position of the V gene in the contig sequence (1-based closed interval).	• VDJ recombination
d_call	First or only D gene segment identified for this contig.	• VDJ recombination
d_support	Quality of D gene alignment, lower is better.	• Sequencing quality • Library quality
d_cigar	CIGAR string for the D gene alignment.	• VDJ recombination
d_sequence_start	Start position of the D gene in the contig sequence (1-based closed interval).	• VDJ recombination

VDJ dominant contigs AIRR output (continued)

Section/metric	Definition	Major contributing factors
d_sequence_end	End position of the D gene in the contig sequence (1-based closed interval).	• VDJ recombination
j_call	J gene segment identified for this contig.	• VDJ recombination
j_support	Quality of J gene alignment - lower is better.	• Sequencing quality • Library quality
j_cigar	CIGAR string for the J gene alignment.	• VDJ recombination
j_sequence_start	Start position of the J gene in the contig sequence (1-based closed interval).	• VDJ recombination
j_sequence_end	End position of the J gene in the contig sequence (1-based closed interval).	• VDJ recombination
c_call	C gene segment identified for this contig.	• VDJ recombination
fwr1	Nucleotide sequence of the FR1 for the contig.	• VDJ recombination
fwr1_aa	Amino acid sequence of the FR1 for the contig.	• VDJ recombination
fwr2	Nucleotide sequence of the FR2 for the contig.	• VDJ recombination
fwr2_aa	Amino acid sequence of the FR2 for the contig.	• VDJ recombination
fwr3	Nucleotide sequence of the FR3 for the contig.	• VDJ recombination
fwr3_aa	Amino acid sequence of the FR3 for the contig.	• VDJ recombination
fwr4	Nucleotide sequence of the FR4 for the contig.	• VDJ recombination
fwr4_aa	Amino acid sequence of the FR4 for the contig.	• VDJ recombination
cdr1	Nucleotide sequence of the CDR1 for the contig.	• VDJ recombination
cdr1_aa	Amino acid sequence of the CDR1 for the contig.	• VDJ recombination
cdr2	Nucleotide sequence of the CDR2 for the contig	• VDJ recombination
cdr2_aa	Amino acid sequence of the CDR2 for the contig.	• VDJ recombination
cdr3	Nucleotide sequence of the CDR3 for the contig.	• VDJ recombination
cdr3_aa	Amino acid sequence of the CDR3 for the contig.	• VDJ recombination

VDJ Unfiltered Contigs (VDJ option selected)

(*runName*)_VDJ_Unfiltered_Contigs_AIRR.tsv

All cell IDs, all assembled contigs that were successfully annotated.

The file is compliant with the AIRR rearrangement schema and contains additional informational columns in addition to all the mandatory ones.

Data columns: Cell Identifiers, Read and Molecule counts, Full trimmed contig nucleotide and amino acid sequence, Framework and CDR region nucleotide and amino acid sequence, V, D, J, and C gene segments, full length, and productive status.

Shared column definitions are identical to the VDJ_Dominant_Contigs_AIRR.tsv file.

Refer to docs.airr-community.org/en/stable/datarep/rearrangements.html#.

VDJ unfiltered contigs AIRR output

Section/metric	Definition	Major contributing factors
sequence	Assembled nucleotide sequence of contig.	• Library quality • VDJ recombination
sequence_length	Length of full contig nucleotide sequence (untrimmed).	• Library quality • VDJ recombination
sequence_aa	Amino acid sequence of contig.	• Library quality • VDJ recombination
sequence_aa_length	Length of full contig amino acid sequence (untrimmed).	• Library quality • VDJ recombination
sequence_alignment	Nucleotide sequence corresponding to VDJ coding region after trimming.	• VDJ recombination
sequence_alignment_length	Length of nucleotide sequence corresponding to VDJ coding region after trimming.	• VDJ recombination
sequence_alignment_aa	Amino acid sequence corresponding to VDJ coding region after trimming.	• VDJ recombination
sequence_alignment_aa_length	Length of amino acid sequence corresponding to VDJ coding region after trimming.	• VDJ recombination
Dominant	True/False—this contig was selected as the dominant contig for this cell-chain combination.	• Library quality • VDJ recombination
Putative_Cell	True/False—this cell index was selected as a putative cell based on the mRNA panel.	• Cell viability • mRNA panel
CDR3_Length	Length of amino acid sequence of the CDR3 for this contig.	• VDJ recombination
fwr1	Nucleotide sequence of the FR1 for the contig.	• VDJ recombination
fwr1_aa	Amino acid sequence of the FR1 for the contig.	• VDJ recombination

VDJ unfiltered contigs AIRR output (continued)

Section/metric	Definition	Major contributing factors
fwr2	Nucleotide sequence of the FR2 for the contig.	• VDJ recombination
fwr2_aa	Amino acid sequence of the FR2 for the contig.	• VDJ recombination
fwr3	Nucleotide sequence of the FR3 for the contig.	• VDJ recombination
fwr3_aa	Amino acid sequence of the FR3 for the contig.	• VDJ recombination
fwr4	Nucleotide sequence of the FR4 for the contig.	• VDJ recombination
fwr4_aa	Amino acid sequence of the FR4 for the contig.	• VDJ recombination
cdr1	Nucleotide sequence of the CDR1 for the contig.	• VDJ recombination
cdr1_aa	Amino acid sequence of the CDR1 for the contig.	• VDJ recombination
cdr2	Nucleotide sequence of the CDR2 for the contig	• VDJ recombination
cdr2_aa	Amino acid sequence of the CDR2 for the contig.	• VDJ recombination
cdr3	Nucleotide sequence of the CDR3 for the contig.	• VDJ recombination
cdr3_aa	Amino acid sequence of the CDR3 for the contig.	• VDJ recombination

Assessing BD Rhapsody™ Analysis pipeline library quality with skim sequencing

Introduction

Several output metrics from the BD Rhapsody™ Analysis pipeline can be evaluated while performing skim sequencing to assess library and sequencing run quality. Output metrics are stable at low sequencing depth (~2 million sequencing reads or higher).

Metrics for evaluation with skim sequencing

Read quality
• Pct_Read_Pair_Overlap • Pct_Reads_Too_Short • Pct_Reads_Low_Base_Quality • Pct_Reads_High_SNF • Pct_Reads_Filtered_Out
Sequencing alignment
• Pct_Q30_Bases_in_Filtered_R2 • Pct_Assigned_to_Cell_Labels • Pct_Cellular_Reads_Aligned_Uniquely
Cells detected
• Putative_Cell_Count (RSEC)[a] • Pct_Reads_from_Putative_Cells (RSEC)[b] • Putative_Cell_Count (DBEC)[a]
a. By metric definition, Putative_Cell_Count (RSEC) has the same value as Putative_Cell_Count (DBEC). Putative_Cell_Count (RSEC) and Putative_Cell_Count (DBEC) might vary by up to ±5% from one sequencing run to the next due to differences in sequencing depth. b. While Pct_Reads_From_Putative_Cells (RSEC) is stable at low sequencing depth, Pct_Reads_From_Putative_Cells (DBEC) is sequencing-depth dependent.

Interpreting output metrics

Introduction

This topic describes possible problems and recommended solutions for sequencing analysis issues. Issues with sequencing metrics might be related to issues that can be resolved in the experimental workflow.

Percentage reads assigned to cell label and percentage cellular reads aligned uniquely to amplicons are low

Possible causes	Recommended solutions
Low sequencing quality	• Ensure that the appropriate PhiX % is used for the type of sequencer used. • Ensure that the Illumina sequencing flow cell is not over-clustered. • Repeat the sequencing run if sequencing quality is suspected to be the reason.
Low library quality	• Ensure that the correct panel is used to amplify the sample and the correct amplification protocol and PCR product purification protocols are used. • Repeat amplification from leftover PCR1 products, if necessary.

High percentage assigned to cell labels but low percentage cellular reads aligned uniquely to amplicons

Possible causes	Recommended solutions
Incorrect FASTA file panel used for mapping	• If <50% alignment, then the wrong panel was likely used. • Verify that the correct panel reference file was used.
Incorrect number of sequencing cycles	Run at least 75 x 2 sequencing cycles. The total length of both reads must be at least 102 bp.
Low sequencing quality	Rerun sequencing, and use at least the minimum recommended concentration of PhiX.

Low percentage reads mapped to putative cells

Possible causes	Recommended solutions
Some cells in the samples are not well represented by the panel. Their associated cell labels have very few detectable molecules, so they are classified as noise cell labels.	• Ensure that the panel matches the sample and species. • Ensure that the panel of genes provides good representation across the cells in the sample tested if all cells are to be detected.
Lysis time too long	Ensure that lysis time is exactly 2 minutes and lysis buffer is cold.
Automated pipette settings are incorrect	Ensure that the correct setting is used for the specific step in the cartridge workflow.
Wrong buffer used for bead retrieval from the cartridge	Use only lysis buffer, as indicated in the protocol for bead retrieval.
Mixed species in experiment	Ensure that the panel used contains genes that cover both species.
Excessive dead or dying cells	Proceed with the experiment if cell viability is ≥50%.
Very low bead loading density. The bead loading efficiency on the BD Rhapsody™ Scanner likely reported failed.	See bead loading density troubleshooting in the *BD Rhapsody™ Single-Cell Analysis System Instrument User Guide* (23-21336) or the *BD Rhapsody™ Express Single-Cell Analysis System Instrument User Guide* (23-21332).

Batch effects across multiple libraries

Possible causes	Recommended solutions
Variations in sequencing depth	Examine the status of each bioproduct in <sample_name>_Bioproduct_Stats.csv across samples. If there are highly abundant genes with a *pass* status in one library but a *low depth* status in another, consider using <sample_name>_RSEC_MolsPerCell.csv for analysis. Or, use <sample_name>_DBEC_MolsPerCell.csv for analysis after removal of genes that do not have *pass* status in any of the libraries under consideration.
Variations in cell sample handling protocol	Use a similar cell sample handling protocol for all samples to be analyzed together, noting that temperature, duration of handling, and handling method can affect bioproduct expression.
Differences in thermal cycling	For samples to be analyzed together, it is recommended to perform the PCR amplification of the Cell Capture Beads of those samples in parallel.
Low sequencing depth	Use <sample_name>_RSEC_MolsPerCell.csv or use <sample_name>_DBEC_MolsPerCell.csv after removal of genes that do not have *pass* status.

Number of cells detected in sequencing is much lower than the expected cell number based on imaging results

Possible causes	Recommended solutions
Some cells in the samples are not well represented by the panel. Their associated cell labels have very few detectable molecules, so they are classified as noise cell labels.	If all of the cells are to be detected, ensure that the panel of genes provides good representation across the cells in the sample tested. Ensure that the panel matches the sample and species. If there is more than one bioproduct type in libraries, use another Putative Cell Calling option to troubleshoot.
Cell Capture Beads settled to the bottom of the tube before the start of PCR1.	Ensure that Cell Capture Beads are well suspended just before starting PCR1, and the thermal cycler lid is pre-heated when the PCR tubes are placed on the thermal cycler.
Cell Capture Beads are lost during handling after cartridge use.	Ensure maximum recovery of Cell Capture Beads by using low retention tips and tubes. See product information in the BD Rhapsody™ *Single-Cell Analysis System Instrument User Guide* (23-21336) or the BD Rhapsody™ *Express Single-Cell Analysis System Instrument User Guide* (23-21332).

References

Bioinformatics analysis tools

- broadinstitute.github.io/picard/. The website contains a set of command line tools for working with high throughput sequencing data and formats, including SAM/BAM/CRAM, and VCF.
- Li H, et al. 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009;25(16):2078–9. doi: 10.1093/bioinformatics/btp352.

- Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nature Methods*. 2012;9(4):357–60. doi: 10.1038/nmeth.1923.
- Fan J, Tsai J, Shum E. Technical Note: Molecular Index counting adjustment methods. BD Biosciences. This is an introduction to RSEC (recursive substitution error correction) and DBEC (distribution-based error correction). For more information, contact BD Biosciences technical support at scomix@bdscomix.bd.com.
- Li H. Toolkit for processing sequences in FASTA/Q formats. github.com/lh3/seqtk.

Expression profiling

Macosko EZ, et al. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell*. 2015;161:1202–1214.

t-distributed stochastic neighbor embedding (t-SNE)

- van der Maaten, LJP. Accelerating t-SNE using Tree-Based Algorithms. *Journal of Machine Learning Research*. 2014; 15(Oct):3221–3245 (PDF).
- van der Maaten LJP, Hinton GE. Visualizing High-Dimensional Data Using t-SNE. *Journal of Machine Learning Research*. 2008; 9(Nov):2579–2605. jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf.

3. Glossary

A

AIRR Adaptive Immune Receptor Repertoire.

B

BAM An alignment file in binary format. A binary SAM file.

Bioproduct Identifiers for biologically-derived products such as mRNA and protein. Examples of identifiers are gene name for mRNA or AbSeq identifier for AbSeq.

Bioproduct Type Type of bioproducts such as mRNA or AbSeq.

C

CIGAR Compact Idiosyncratic Gapped Alignment Report. A sequence of base lengths to indicate base alignments, insertions, and deletions with respect to the reference sequence. See samtools.github.io/hts-specs/SAMv1.pdf.

CLS Cell label sequence.

D

DBEC Distribution-based error correction.

F

FASTA Text-based format that contains one or more DNA or RNA sequences.

FASTQ A file in standardized, text-based format that contains the

	output of read bases and per-base quality values from a sequencer.

L

L	Common sequence.

M

molecule	A unique combination of a cell label, UMI sequence, and a bioproduct. Without UMI adjustment methods, it is called *raw molecule*. With RSEC UMI adjustment, it is called *RSEC-adjusted molecule*. With additional DBEC UMI adjustment, it is called *DBEC-adjusted molecule*.

P

PhiX	Control library used for sequencing runs.

R

R1 reads	Contains information about the cell label and UMI.
R2 reads	Contains information about the bioproduct.
RSEC	Recursive substitution error correction.

S

SAM	Tab-delimited text file with sequence alignment data.
singlet	A putative cell where more that 75% of sample tag reads are from a single tag.
singleton	Clustering: Cell not assigned to any of the clusters. UMI correction/adjustment: Molecule that is represented by only one

read.

U

UMI Unique Molecular Identifier. A string of eight randomers immediately downstream of the cell label sequence (CLS) 3 of the R1 read that is used to uniquely label a molecule.

Becton, Dickinson and Company
BD Biosciences
2350 Qume Drive
San Jose, California 95131 USA

bdbiosciences.com
scomix@bdscomix.bd.com